Software Intelligence with

Codeface

Understanding the Linux Ecosystem

Dr. Wolfgang Mauerer, M. Joblin, Dr. J. Ebke,
Dr. M. Meilinger, Dr. A. Eckert, S. Weber, Prof. Dr. S. Apel,
C. Rubner, D. Srivastav

Siemens AG, Corporate Research and Technologies
Corporate Competence Centre Embedded Linux
`wolfgang.mauerer@siemens.com`

OTH Ostbayerische Technische Hochschule Regensburg
Universität Passau

SIEMENS

Overview

What's Open Source Software about?

- Publishing code
- Building communities → *Network effects!*

So there are network effects…BUT:

- Different types?
- Good and bad approaches?
- Impact of technology?
- Better software?

SIEMENS

Overview

What's Open Source Software about?

- Publishing code
- Building communities → *Network effects!*

So there are network effects…BUT:

- Different types?
- Good and bad approaches?
- Impact of technology?
- Better software?

SIEMENS

Overview II

Why?

- Learn about *embedded software ecosystem*
- Find proper *involvement*
- Project *controlling*
- Learn from masters!

The two problems of software development

SIEMENS

Problem 1: It's about Technology

SIEMENS

Problem 2: It's about People

Project: Software Intelligence

Pharmaceuticals

- ✓ A-priori understanding
- ✓ Tests & statistics

Image source: `fraxinusit.com`

Project: Software Intelligence

Pharmaceuticals

- ✓ A-priori understanding
- ✓ Tests & statistics

Software

- ✗ Comparative experiments
- ✗ Quantify people and behaviour
- ✗ Personal experience limited

Image source: `fraxinusit.com`

Introducing Codeface

Codeface

Codeface: Two Aspects

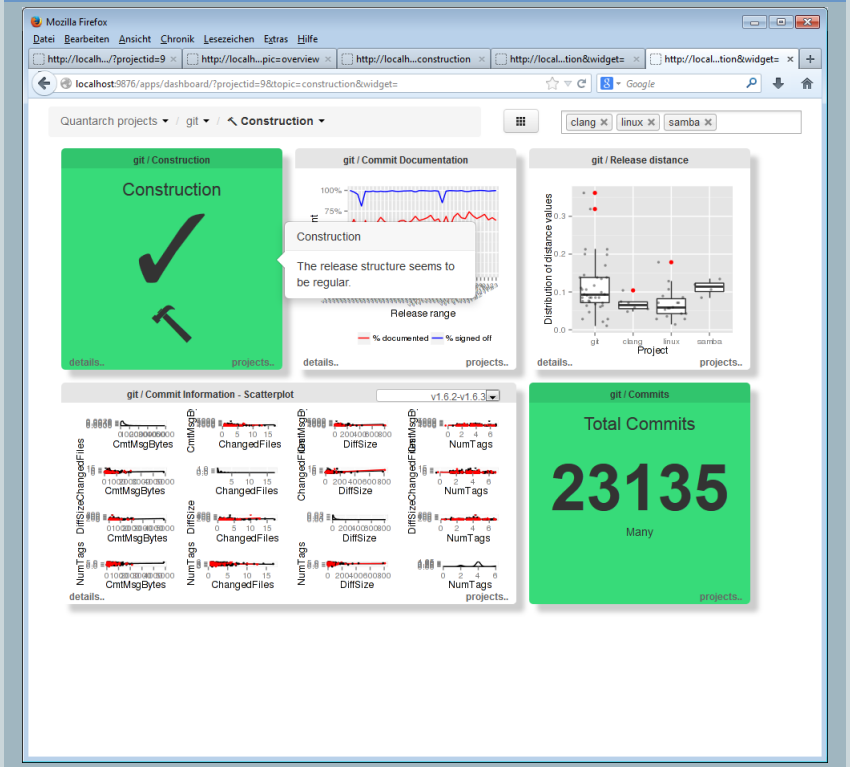
(Fundamental) Research

$$P_{R_i}(k) = P_G(k) \; \forall R_i \in \mathcal{R}$$

$$P_{\mathcal{R}}(Q) = \frac{|\{i \in [1, |\mathcal{R}|] \mid q_{R_i}(\mathcal{C}) = Q\}|}{|\mathcal{R}|}$$

$$H_0 : P_{\mathcal{R}}(Q = q_G(\mathcal{C})) > \epsilon$$

Practical Software



SIEMENS

Project Goals

Learn from SW Devel Data

- Objective properties of successful projects
- Understand network effects
- Quantify social factors
- Mine large bodies of software (calibration: open source)

Find *Actionable* Insights

- Find most efficient approaches *for a given scenario*
- Assess ongoing development
- Make informed choices

-ENOINTENTION

Don't impose mandatory interpretations!

SIEMENS

Project Goals

Learn from SW Devel Data

- Objective properties of successful projects
- Understand network effects
- Quantify social factors
- Mine large bodies of software (calibration: open source)

Find *Actionable* Insights

- Find most efficient approaches *for a given scenario*
- Assess ongoing development
- Make informed choices

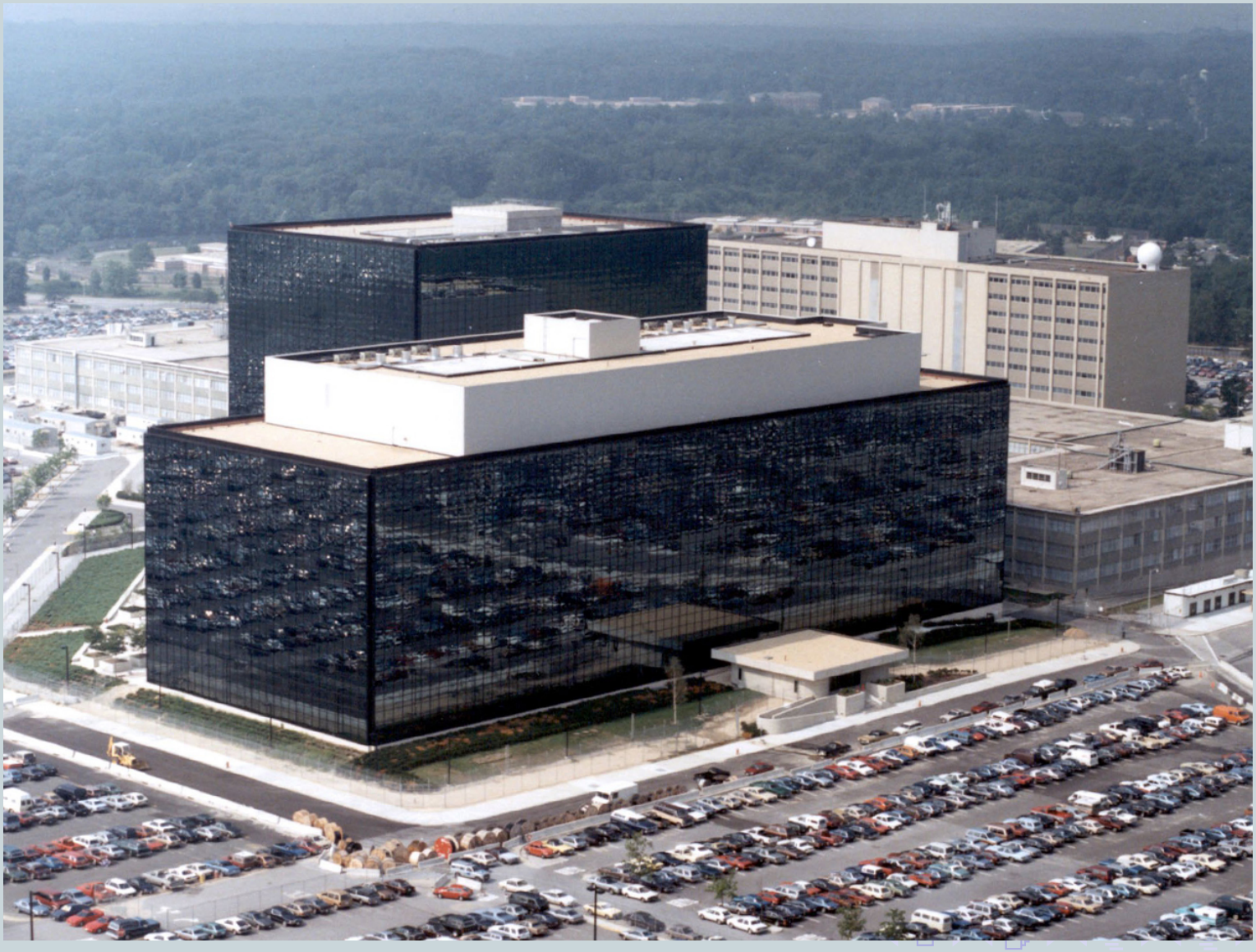
-ENOINTENTION

Don't impose mandatory interpretations!

Data Sources

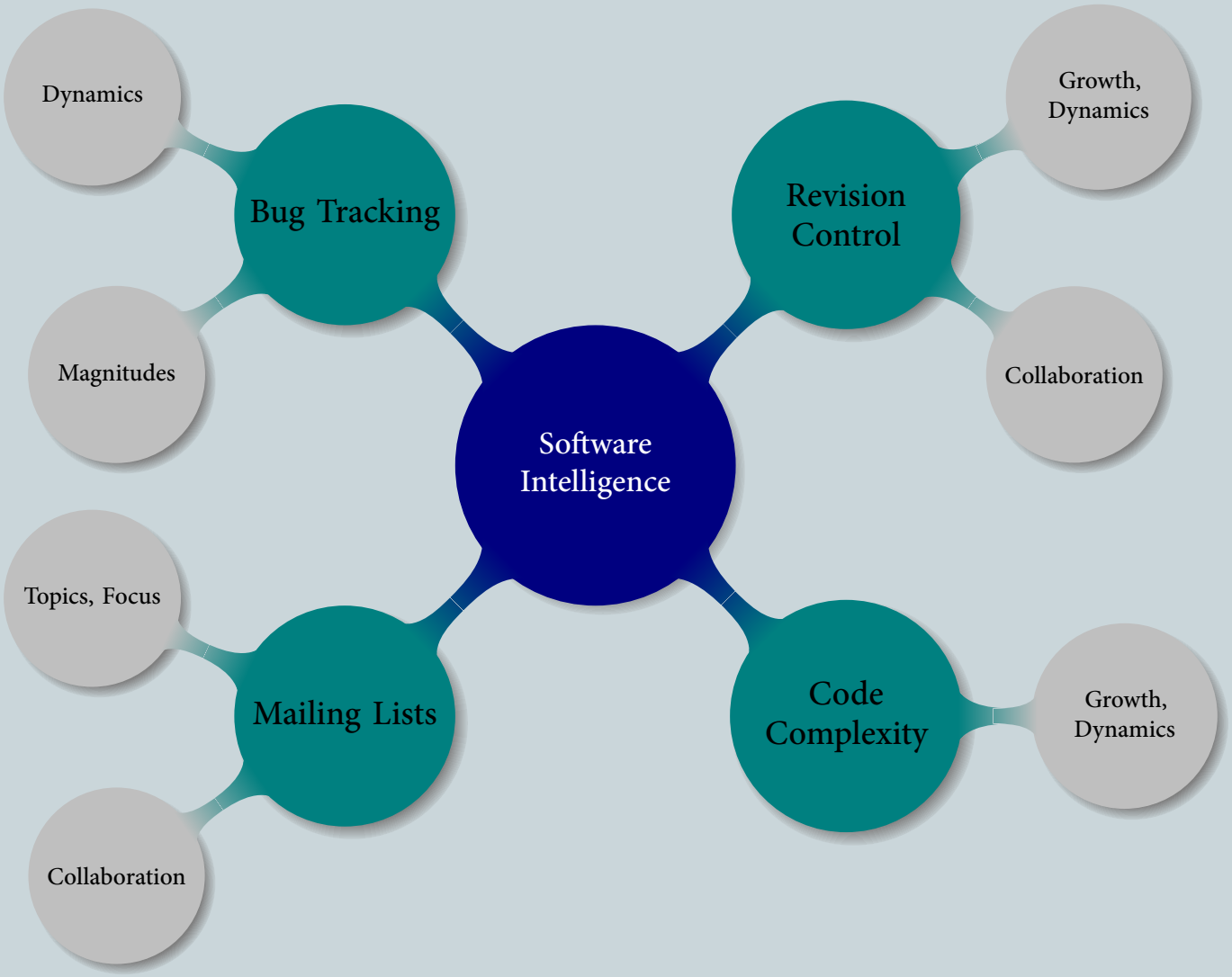
SIEMENS

Data Sources



SIEMENS

Data Sources



Examples

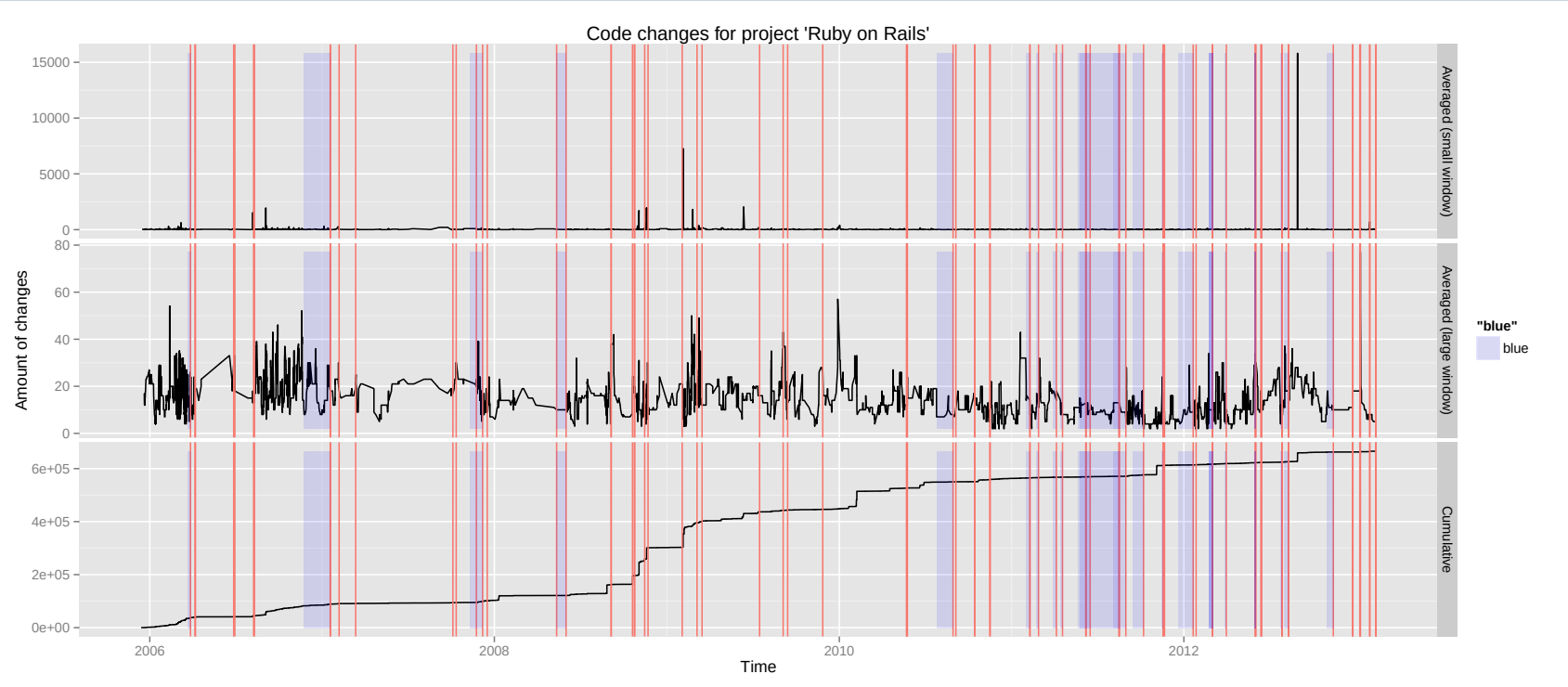
SIEMENS

Example – Time Series

Example – Time Series

SIEMENS

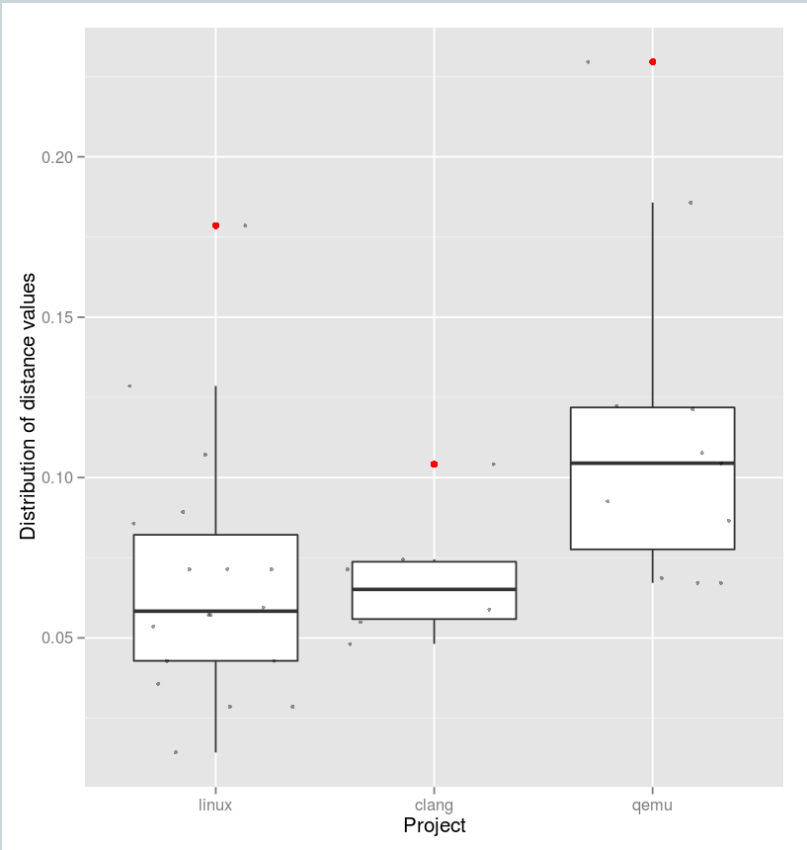
Example – Time Series



Time Series II

Interpretation

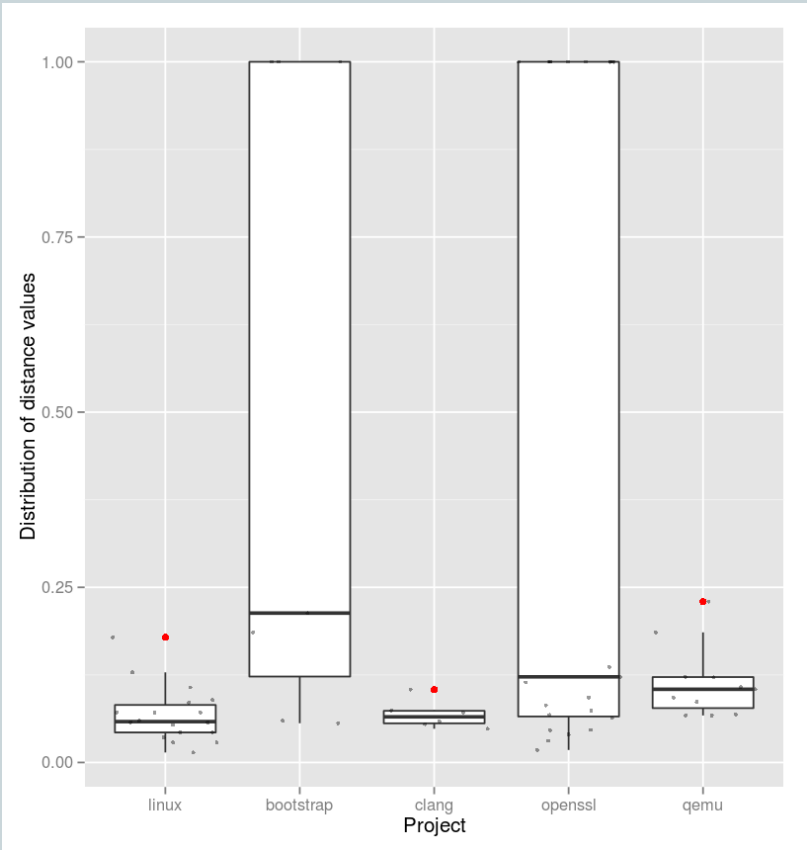
- No objectively „optimal“ approach
- Self-Consistency matters!



Time Series II

Interpretation

- No objectively „optimal“ approach
- Self-Consistency matters!

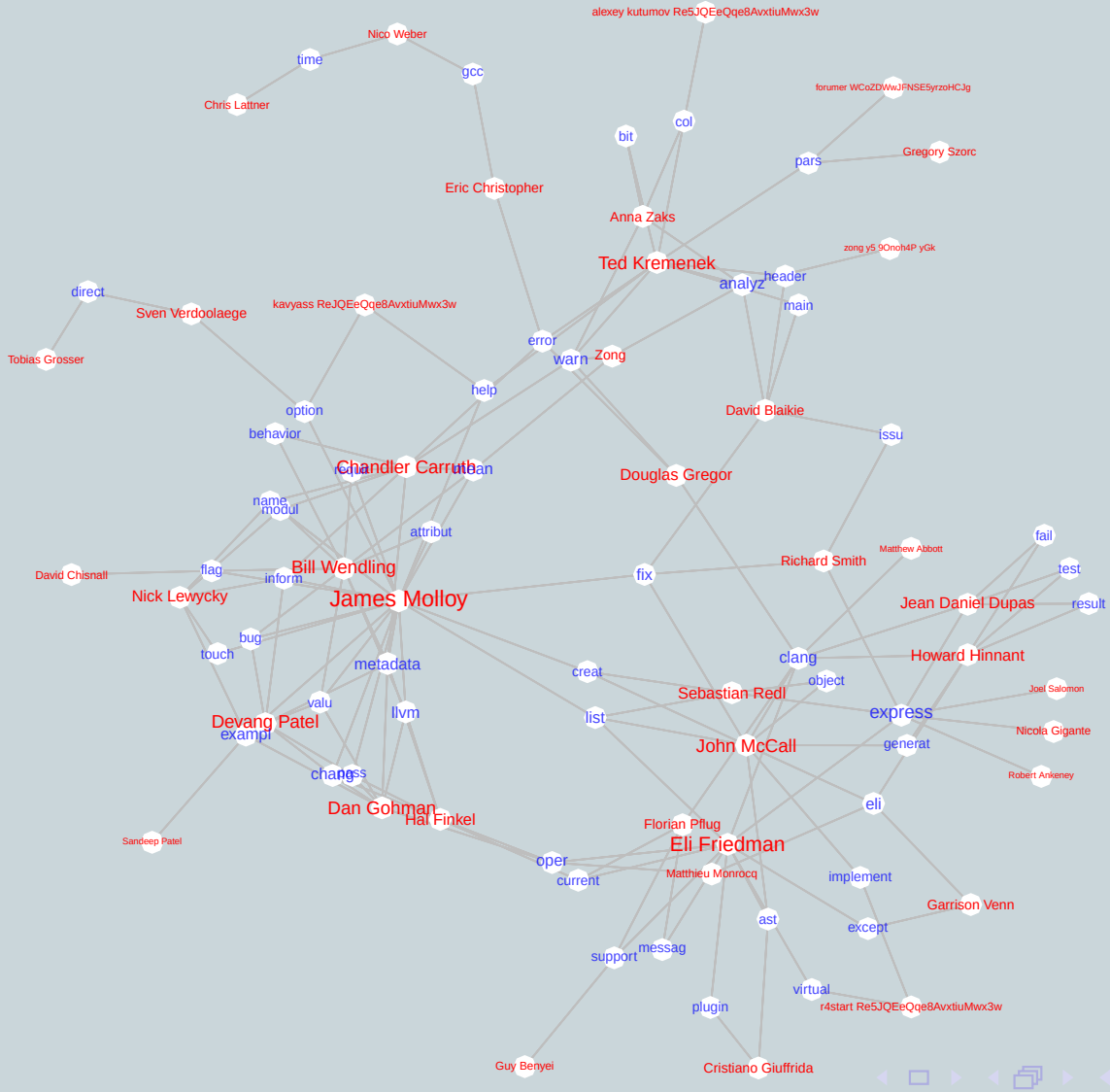


SIEMENS

Example – Communication: Paths & Topics

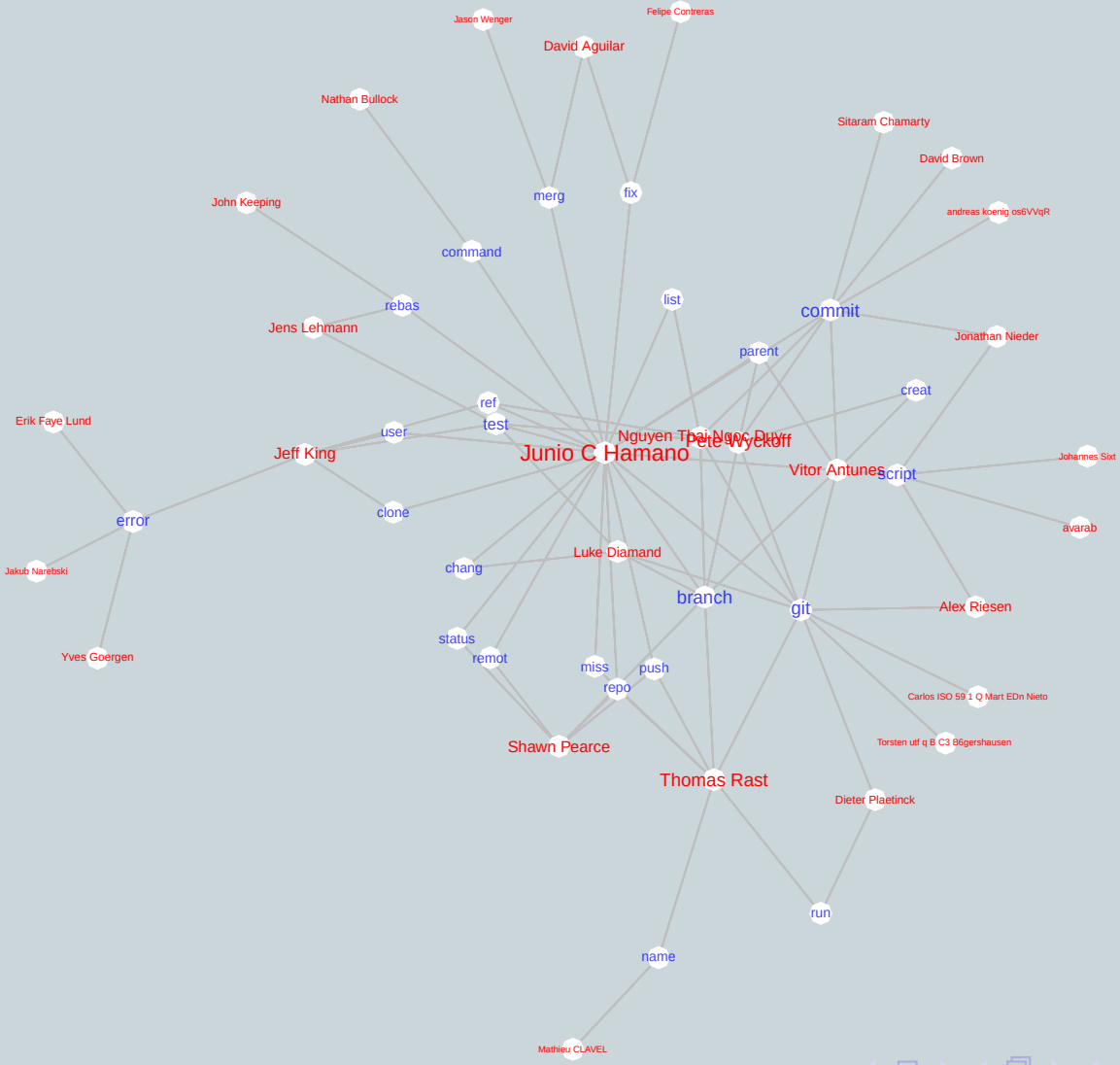
SIEMENS

Example – Communication: Paths & Topics



SIEMENS

Example – Communication: Paths & Topics



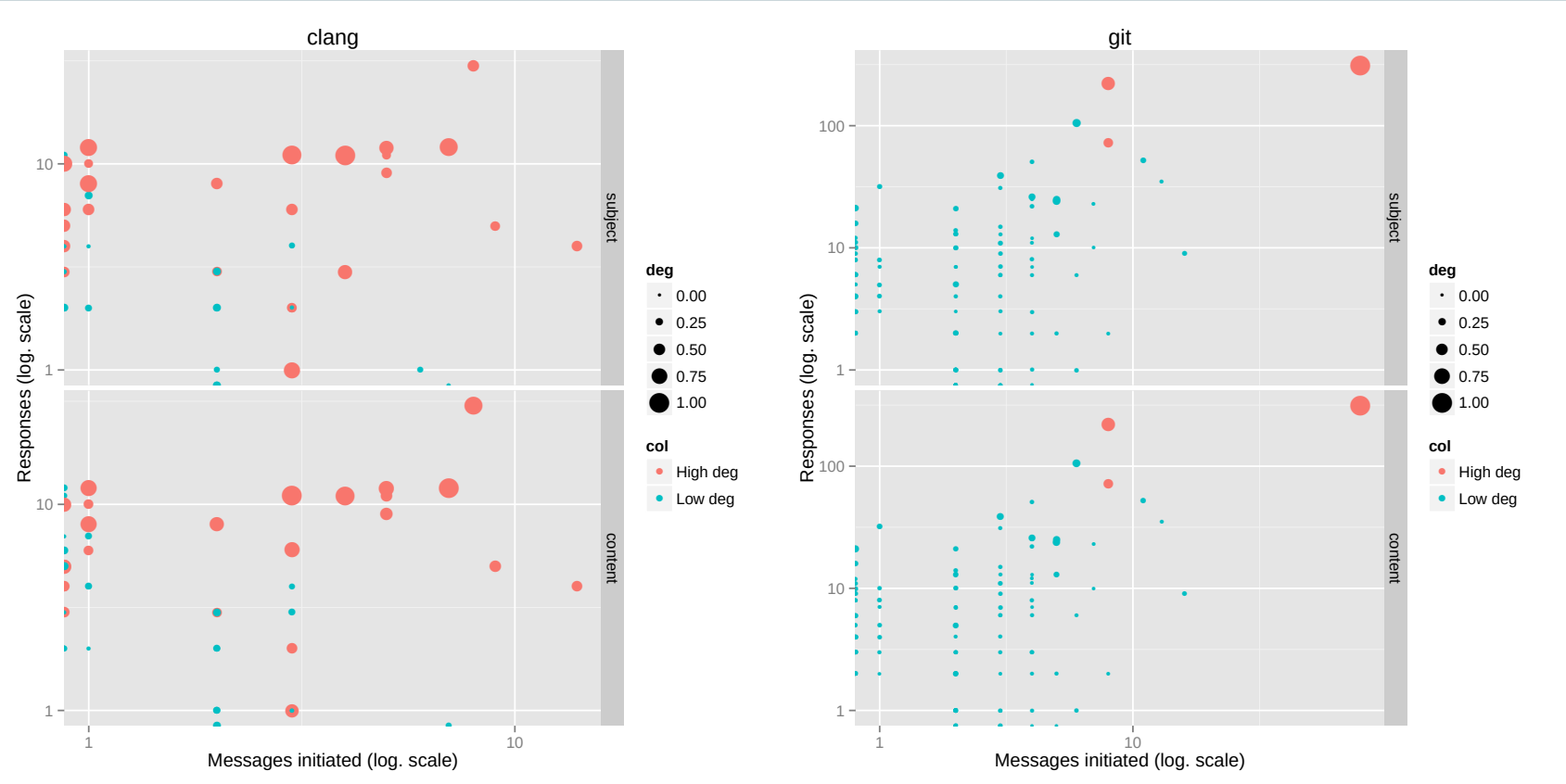
SIEMENS

Communication Paths & Topics

It's not the content…

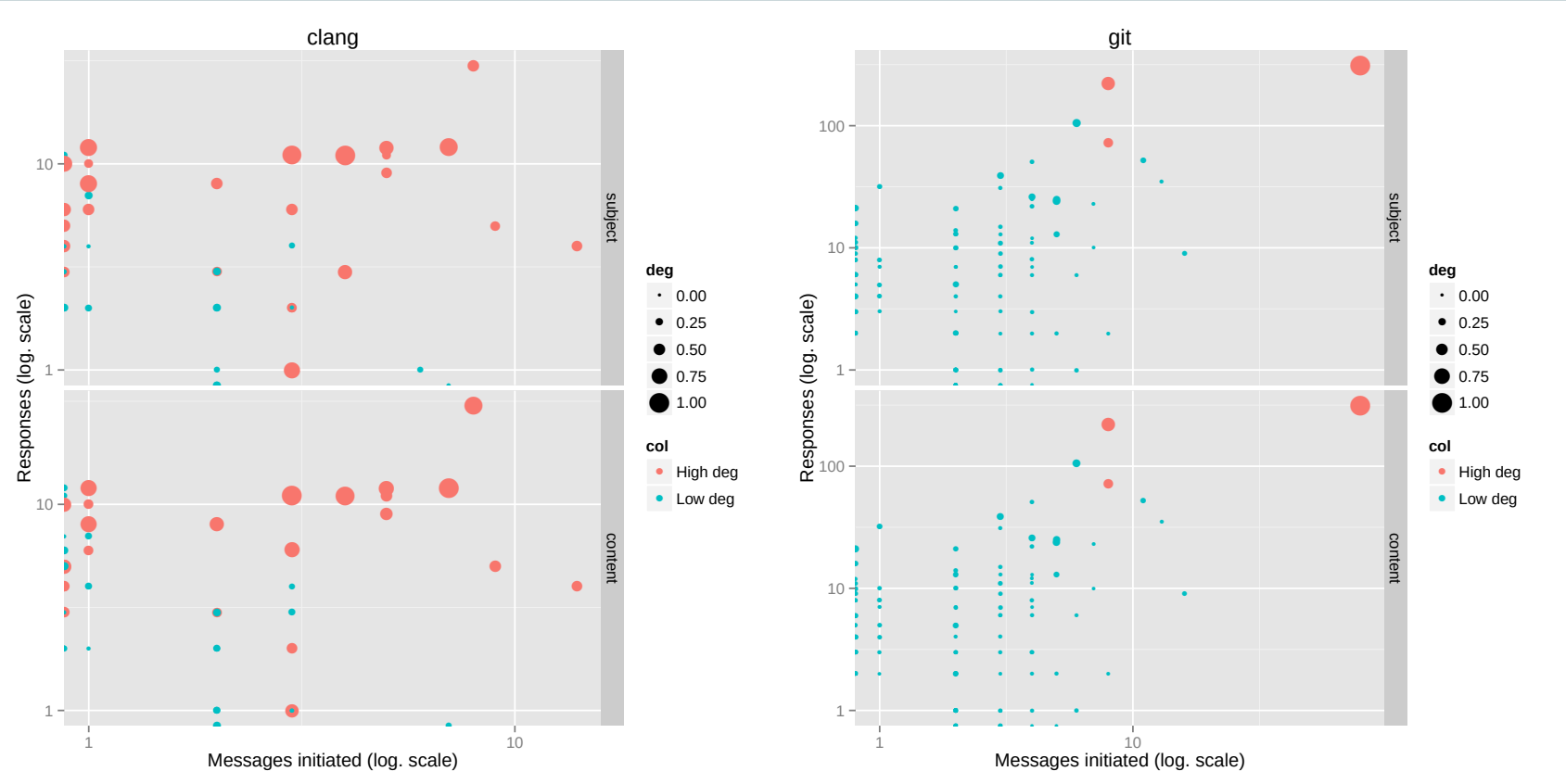
- Topics: No surprises
- Hard to evaluate by machines
- *But:* Overall picture meaningful

Examples – Communication Paths & Topics



Content does not matter—subjects suffice!

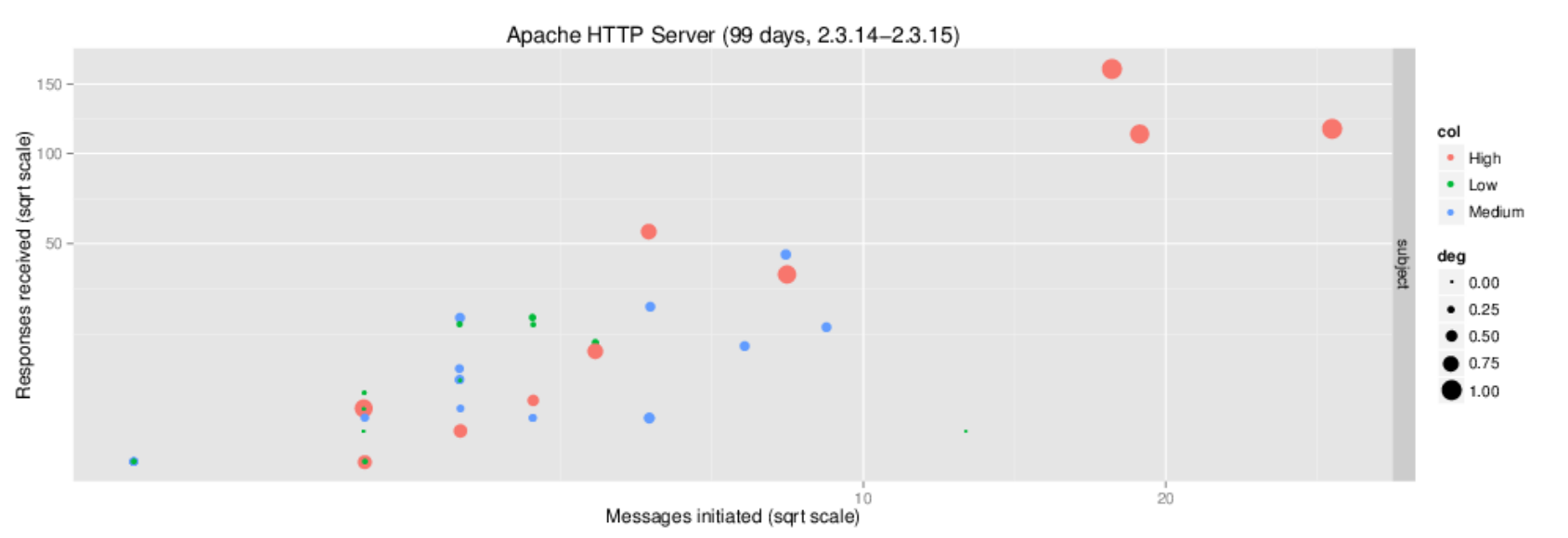
Examples – Communication Paths & Topics



Content does not matter—subjects suffice!

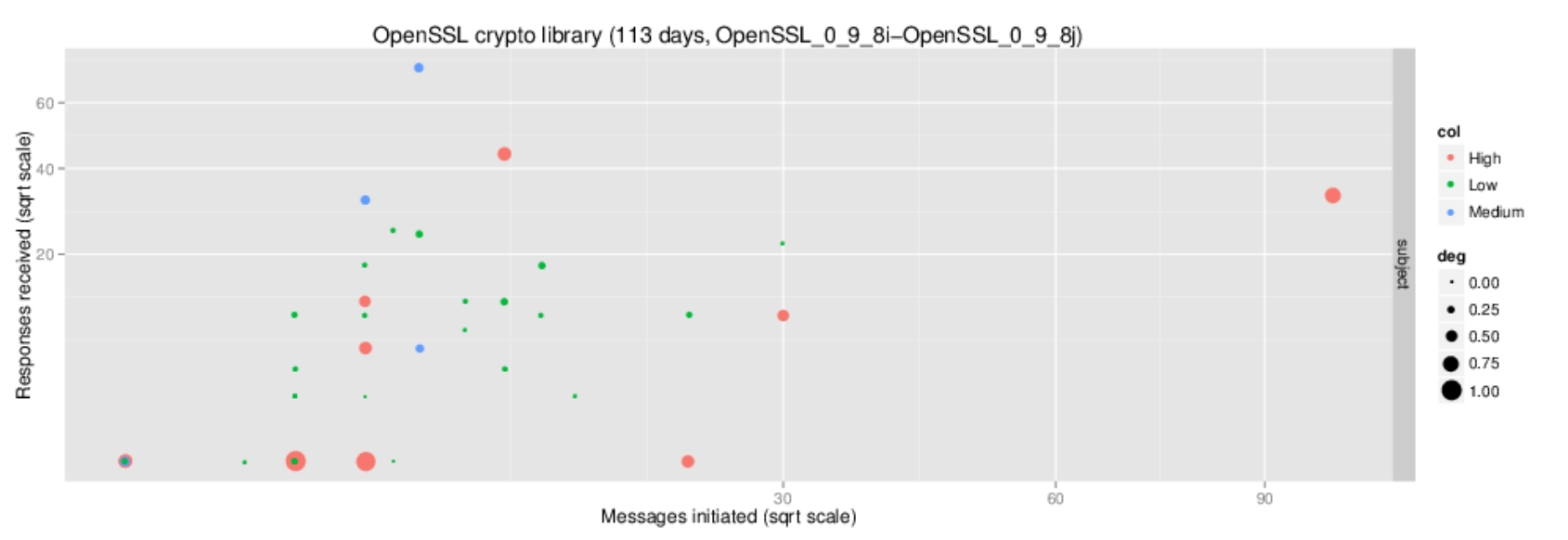
SIEMENS

Communication Paths & Topics II



SIEMENS

Communication Paths & Topics II

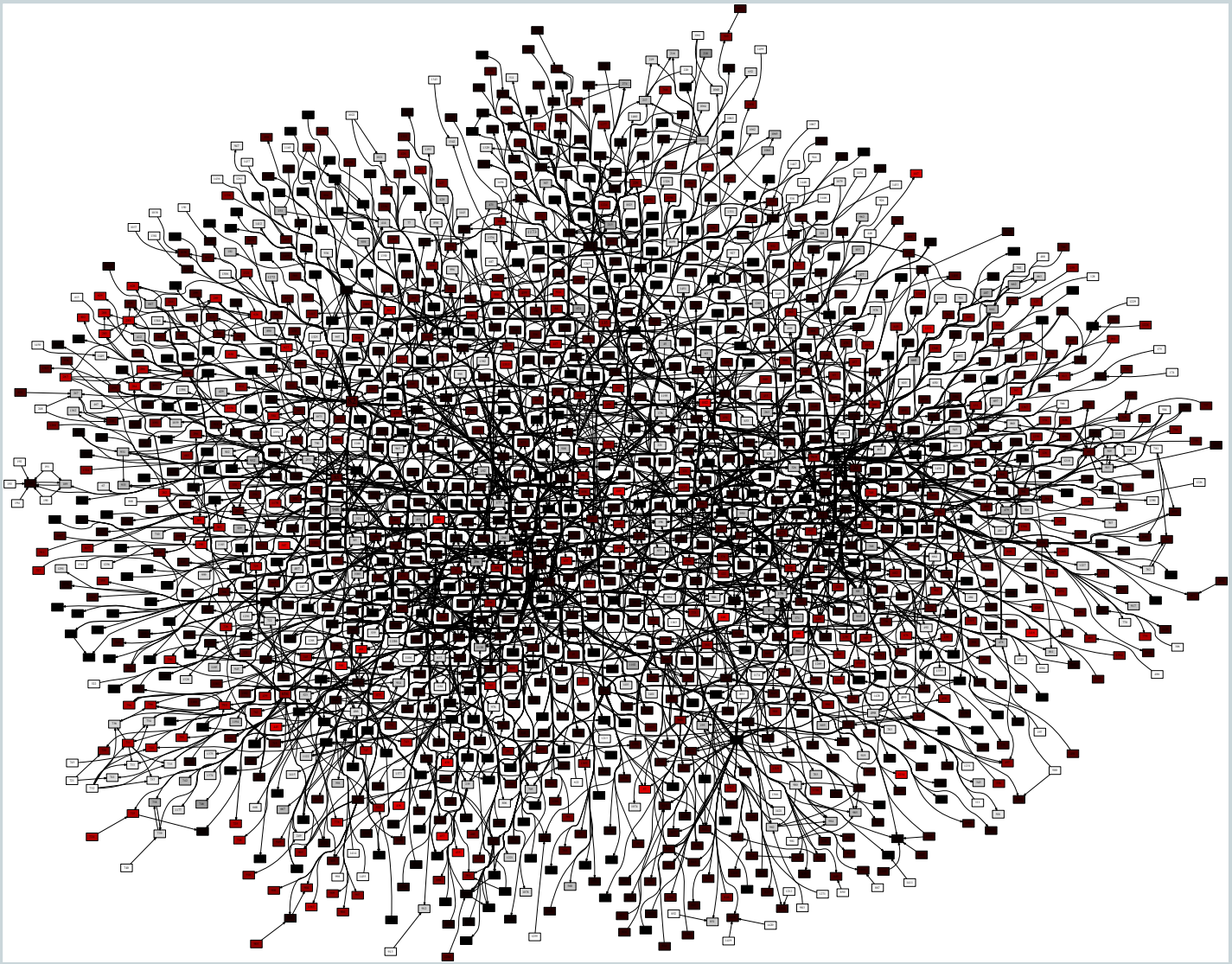


SIEMENS

Examples – Cooperation and Teams

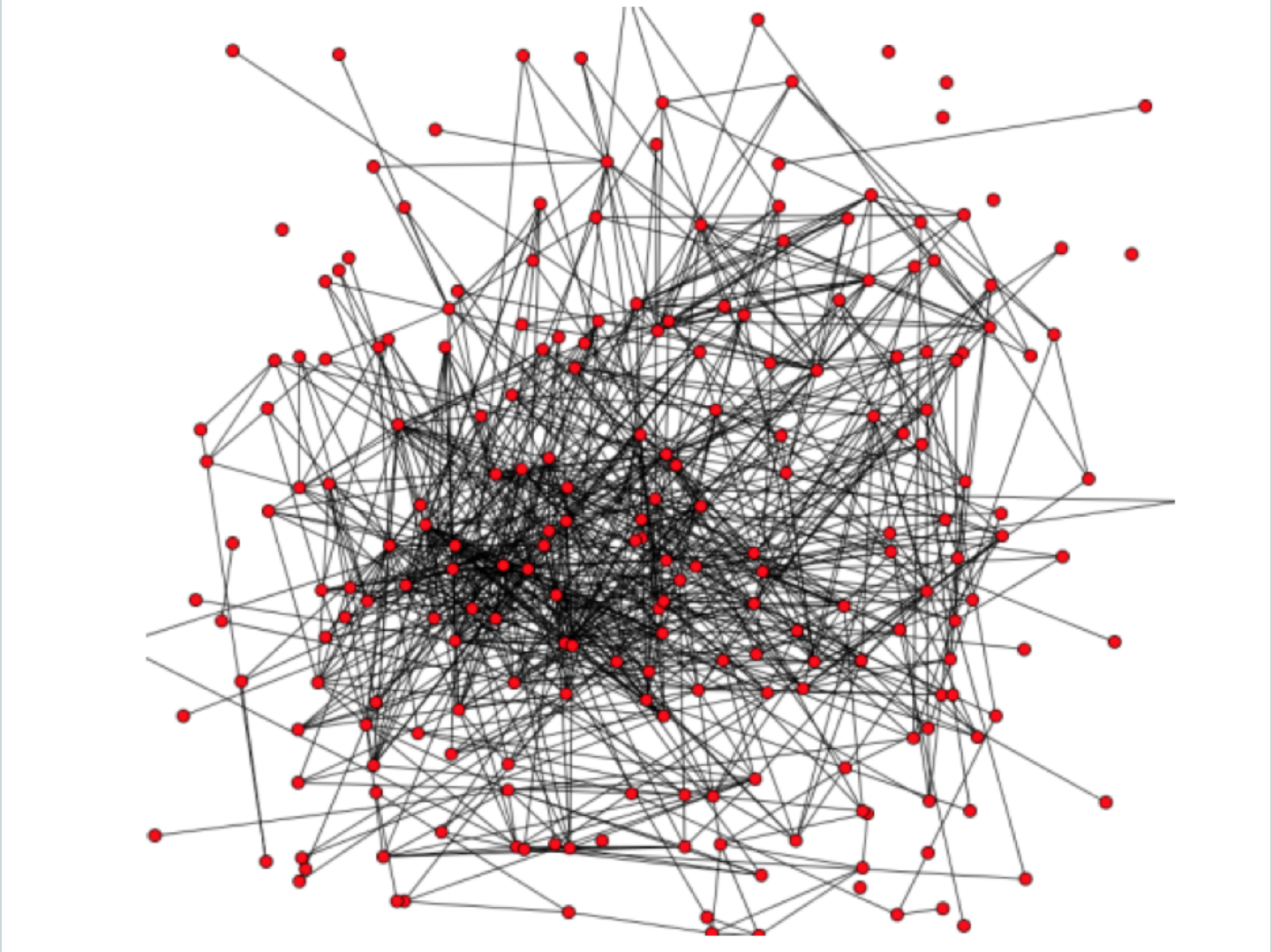
SIEMENS

Examples – Cooperation and Teams



SIEMENS

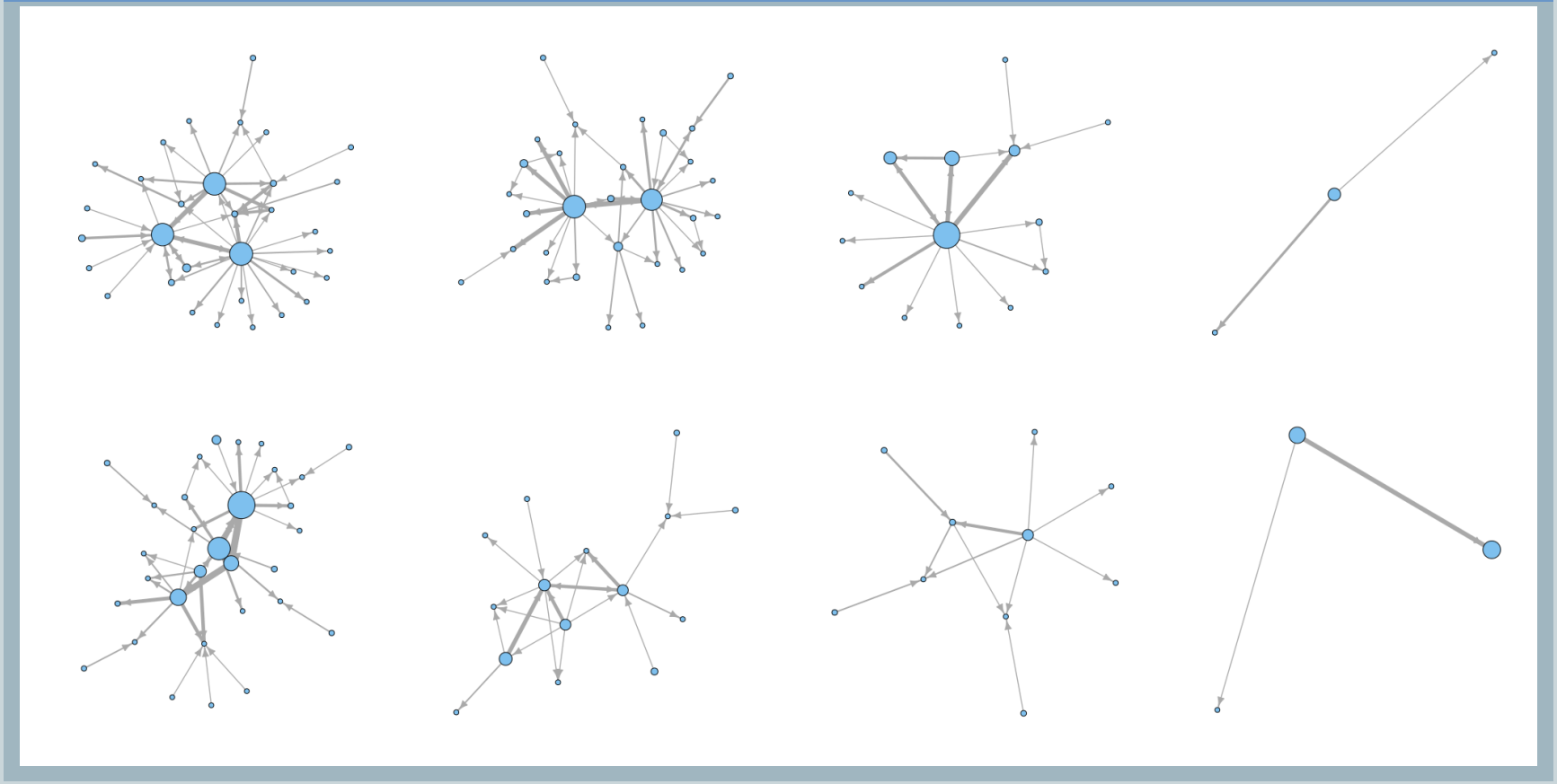
Examples – Cooperation and Teams



SIEMENS

Cooperation I

QEMU: Current State

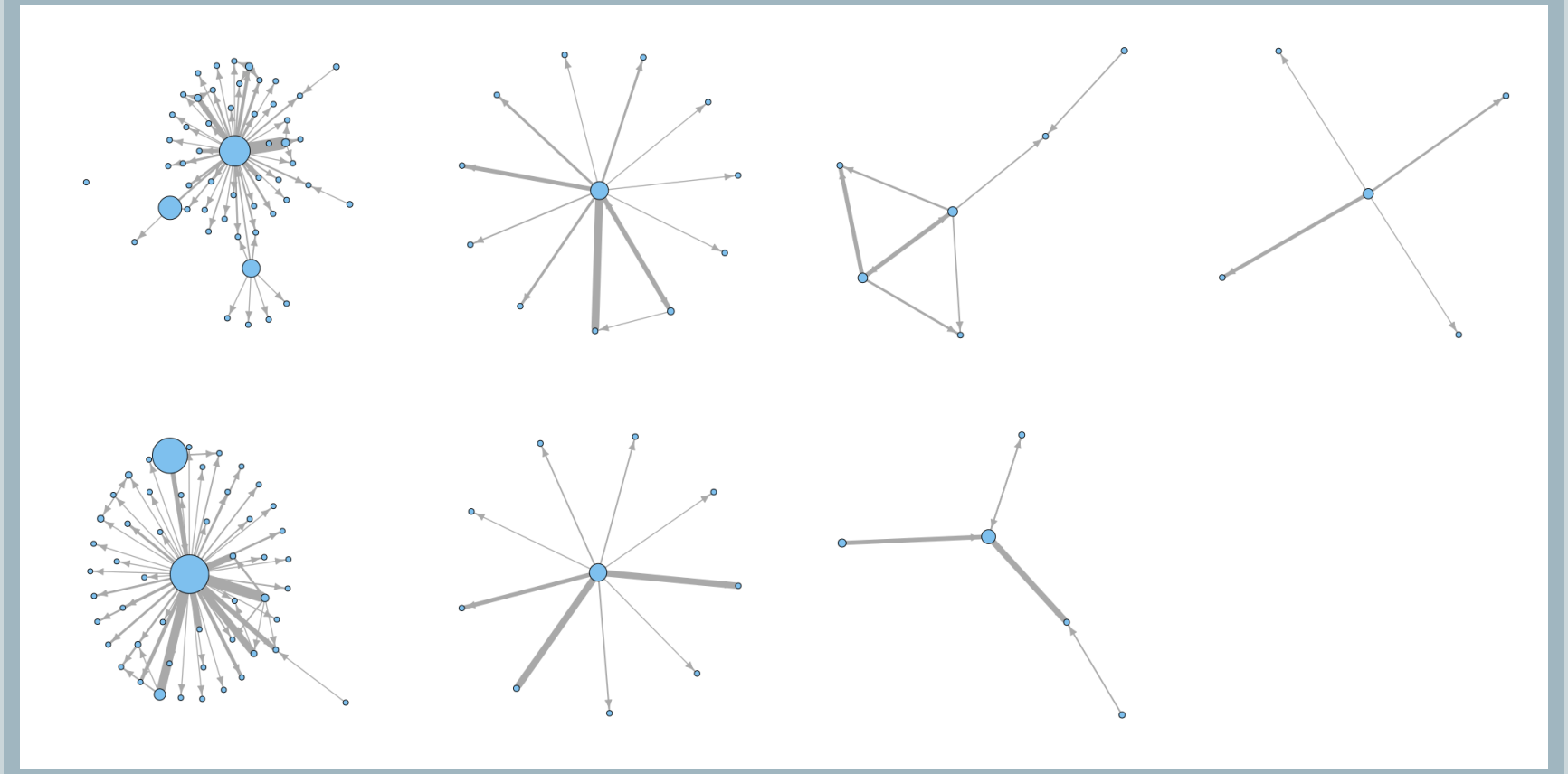


- Influence of individual contributors
- Group structure

SIEMENS

Cooperation I

QEMU: Early Days



- Influence of individual contributors
- Group structure

SIEMENS

Cooperation II

How to…

- Construct network?
- Determine contributor centrality?
- Identify (meaningful) subgroups?

Cooperation III

Network construction

- Tagging (“Signed-off-by”)
- Committer-Author
- Overlapping contributions

Contributor Centrality

Google’s Page Rank

Community Decomposition

- Spin Glass approach

Image source: `cdn.oliveandcocoa.com/`

Cooperation III

Network construction

- Tagging (“Signed-off-by”)
- Committer-Author
- Overlapping contributions

Contributor Centrality

Google’s Page Rank

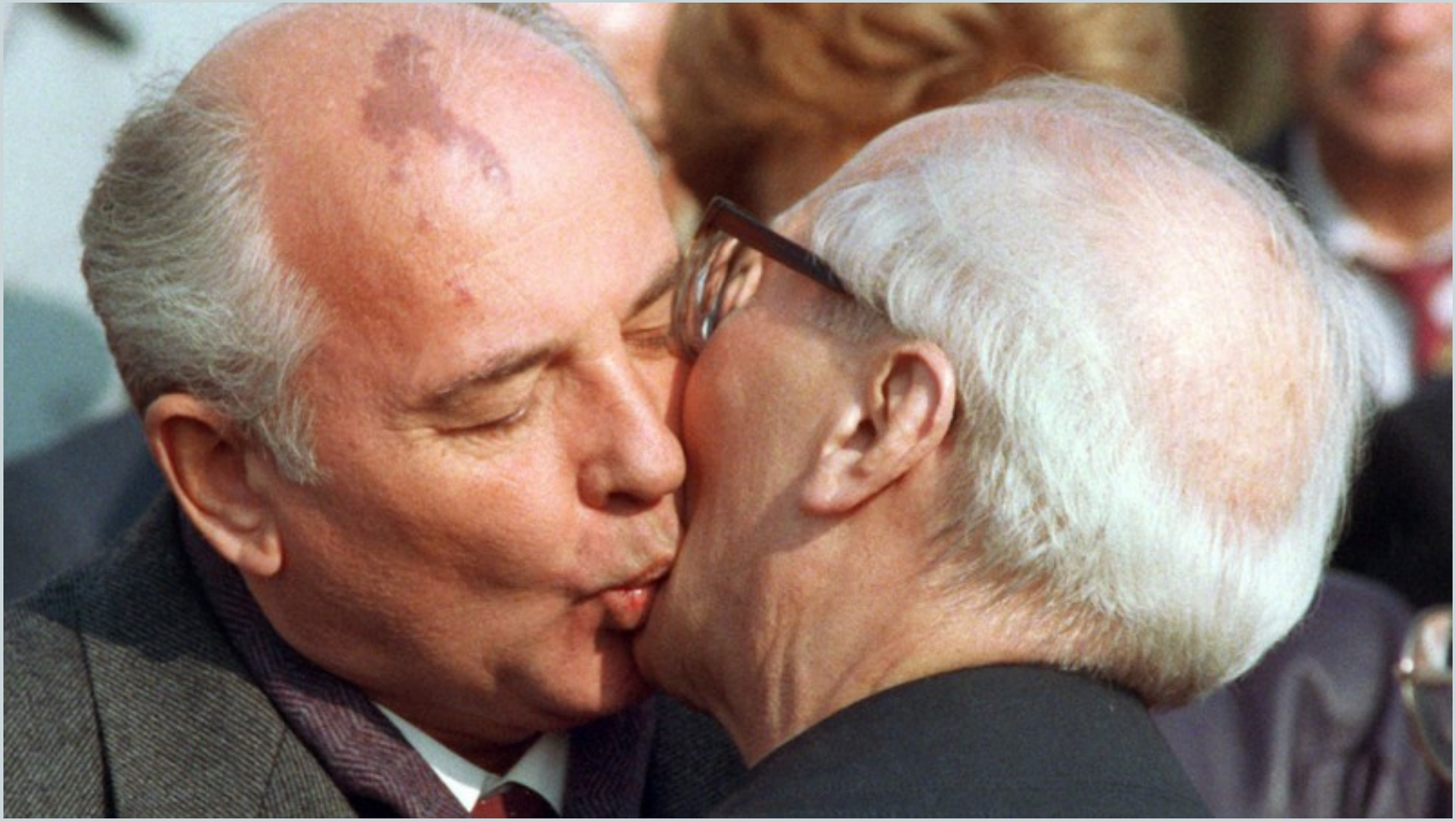
Community Decomposition

- Spin Glass approach
- Vertices (developers) have spin state $\in [0, c]$
- Edges (connections): Preferences to spin states
- Approximate lowest energy state
- Multi-Method support: OSLOM, Random Walks

Image source: `cdn.oliveandcocoa.com/`

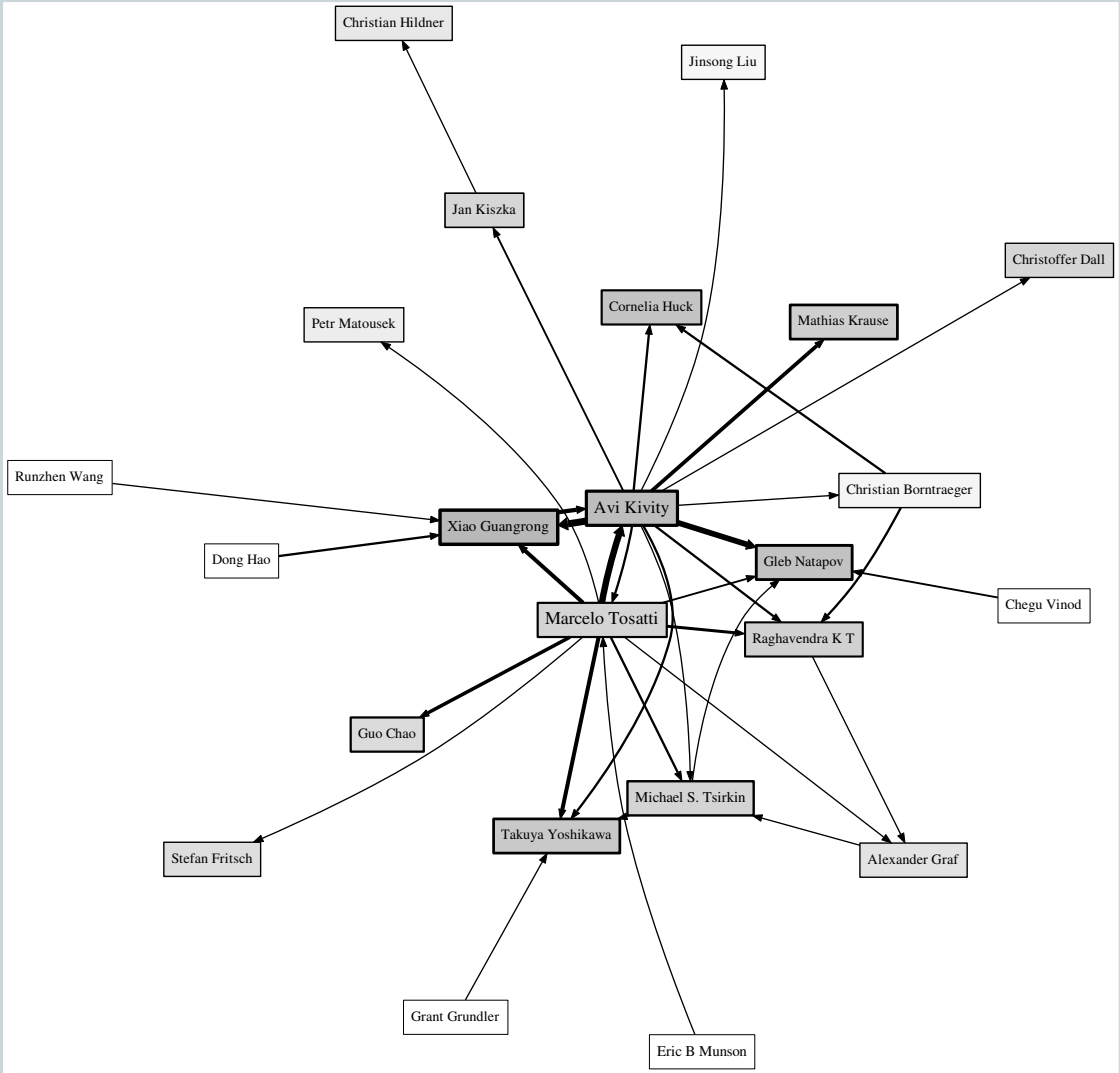
SIEMENS

It's about People, part 2: Trust

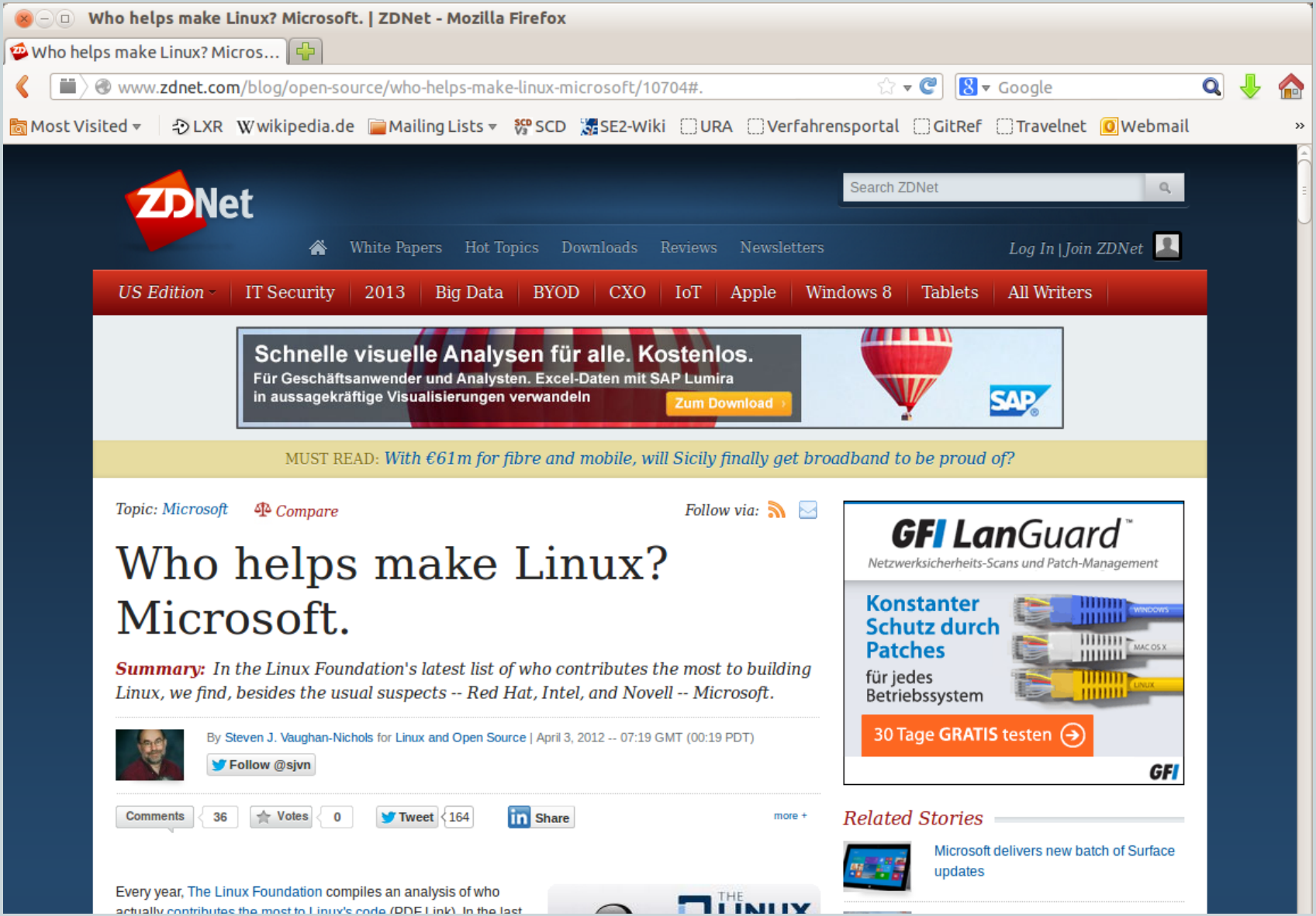


SIEMENS

Community Validation: Expert Knowledge



Microsoft is taking over the Linux kernel!



SIEMENS

Microsoft is taking over the Linux kernel!

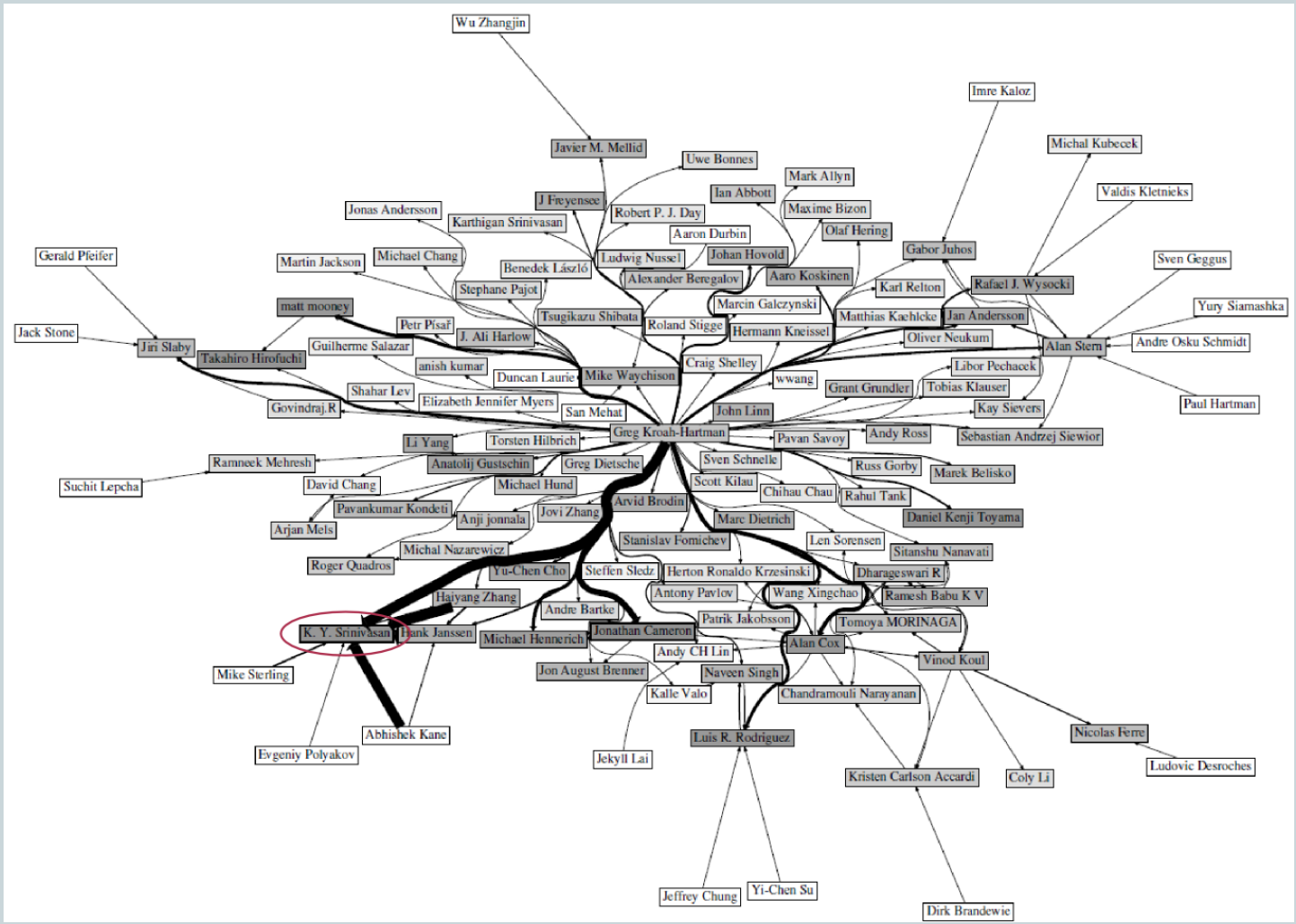
Top Linux 3.0 Contributors (# Commits)

K. Y. Srinivasan	343	3.8%
David S. Miller	176	2.0%
Dan Williams	149	1.7%
Jonathan Cameron	119	1.3%
Takashi Iwai	108	1.2%
Mark Brown	91	1.0%

- Surely people $\neq f(x_1, x_2, \ldots, x_n)$
- Maybe persons $\approx f(x_1, x_2, \ldots, x_n | \text{large context})$

SIEMENS

Communities vs. Numbers



SIEMENS

Let's go Quantitative

Nonsense

- Black xor White
- Optimal approach
- Counting apples and peas per developer per minute times square foot

Sense

- How does $\langle A \rangle$ compare to $\langle B \rangle$?
- Self-consistency
- Likelihood of result compared to random approach

Community Decomposition and Quality Estimation I

Quality Estimation

- Meaningful community structures vs. random properties
- Randomise clusters
  - Rewire edges
  - Keep properties (e.g., “amount” of participation)
- H_0: Clustering stems from unorganised, random process.
- Reject $\rightarrow$ Decomposition makes sense

Alternative Approach

Ken Schwaber says: A team has seven people (plus or minus two). Full stop!

Community Decomposition and Quality Estimation I

So let's study the Maths!

Conductance $\phi \in [0\,,\,1]$ of a community C that is a sub-graph of a larger collaboration graph G is defined as

$$\phi_G(C) := \frac{|(C, G\ C)|}{\{(C), (G\ C)\}}\,, \tag{1}$$

Identify a set of communities $\mathcal{C} = \{C_1\,,\,C_2\,,\,\ldots\,,\,C_i\}$, $C_i \subseteq G$. Conductance over all communities is given by $q_G(\mathcal{C}) = \sum_{C\in\mathcal{C}} \phi_G(C)/|\mathcal{C}|$. Rewire an edge pair $(u\,,\,v)$ and $(s\,,\,t)$ is rewired to $(u\,,\,t)$ and $(s\,,\,v)$, maintaining the number of edges) but destroying the preferential attachement. Repeat to generate a set $\mathcal{R} = \{R_1\,,\,R_2\,,\,\ldots\,,\,R_m\}$ of randomized graphs with $V(R_i) = V(G)\ \forall i$. $P_C(k) = |\{c \in C|\,(c) = k\}|/|N|$ ist maintained, so

$$P_{R_i}(k) = P_G(k)\ \forall R_i \in \mathcal{R}\,. \tag{2}$$

Define

$$P_{\mathcal{R}}(Q) = \frac{|\{i \in [1\,,\,|\mathcal{R}|]\,|\,q_{R_i}(\mathcal{C}) = Q\}|}{|\mathcal{R}|}\,. \tag{3}$$

and test the hypothesis

$$H_0 : P_{\mathcal{R}}(Q = q_G(\mathcal{C})) > \epsilon\,, \tag{4}$$

against the alternative hypothesis given by $H_1 : P_{\mathcal{R}}(\cdot) \leq \epsilon$.

Community Decomposition and Quality Estimation I

Community Decomposition and Quality Estimation I

Quality Estimation

- Meaningful community structures vs. random properties
- Randomise clusters
  - Rewire edges
  - Keep properties (e.g., “amount” of participation)
- H_0: Clustering stems from unorganised, random process.
- Reject $\rightarrow$ Decomposition makes sense

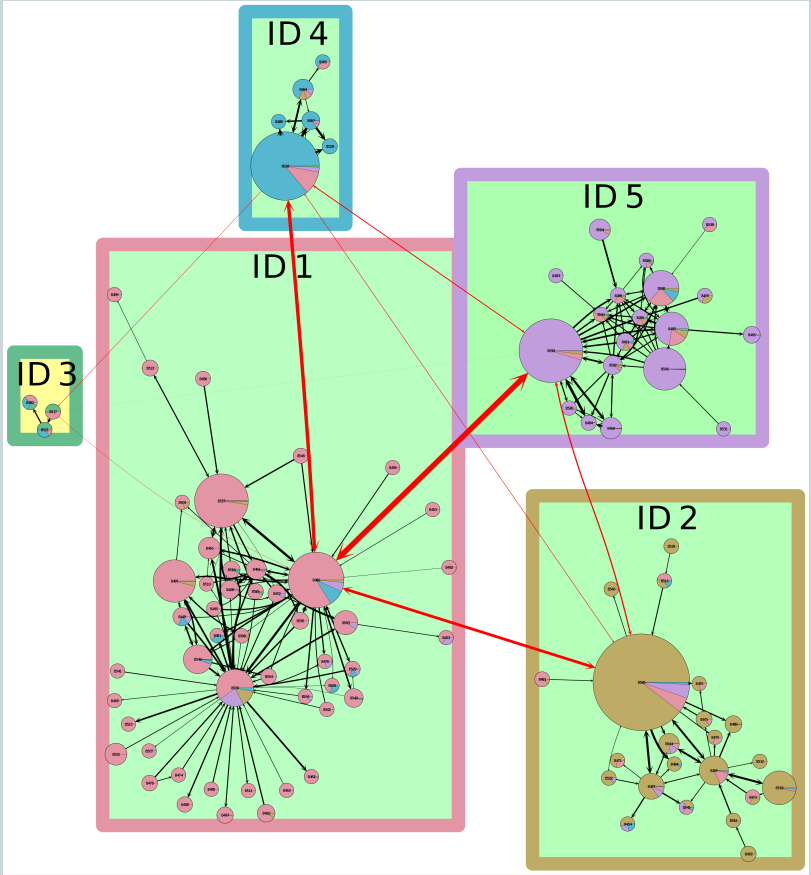
Alternative Approach

Ken Schwaber says: A team has seven people (plus or minus two). Full stop!

Community Decomposition and Quality Estimation III

Cheat Sheet

- Boxes: developer communities
- Background color: strength of community (green=strong, yellow=weak)
- Box border color: unique community identifier
- Pie chart: fraction of developer participation in given community color
- Node size: developers importance according to centrality measure
- Link thickness: strength of relationship between entities
- Link color: red is inter-community



SIEMENS

Input needed: Validate/Refute

Developer Input

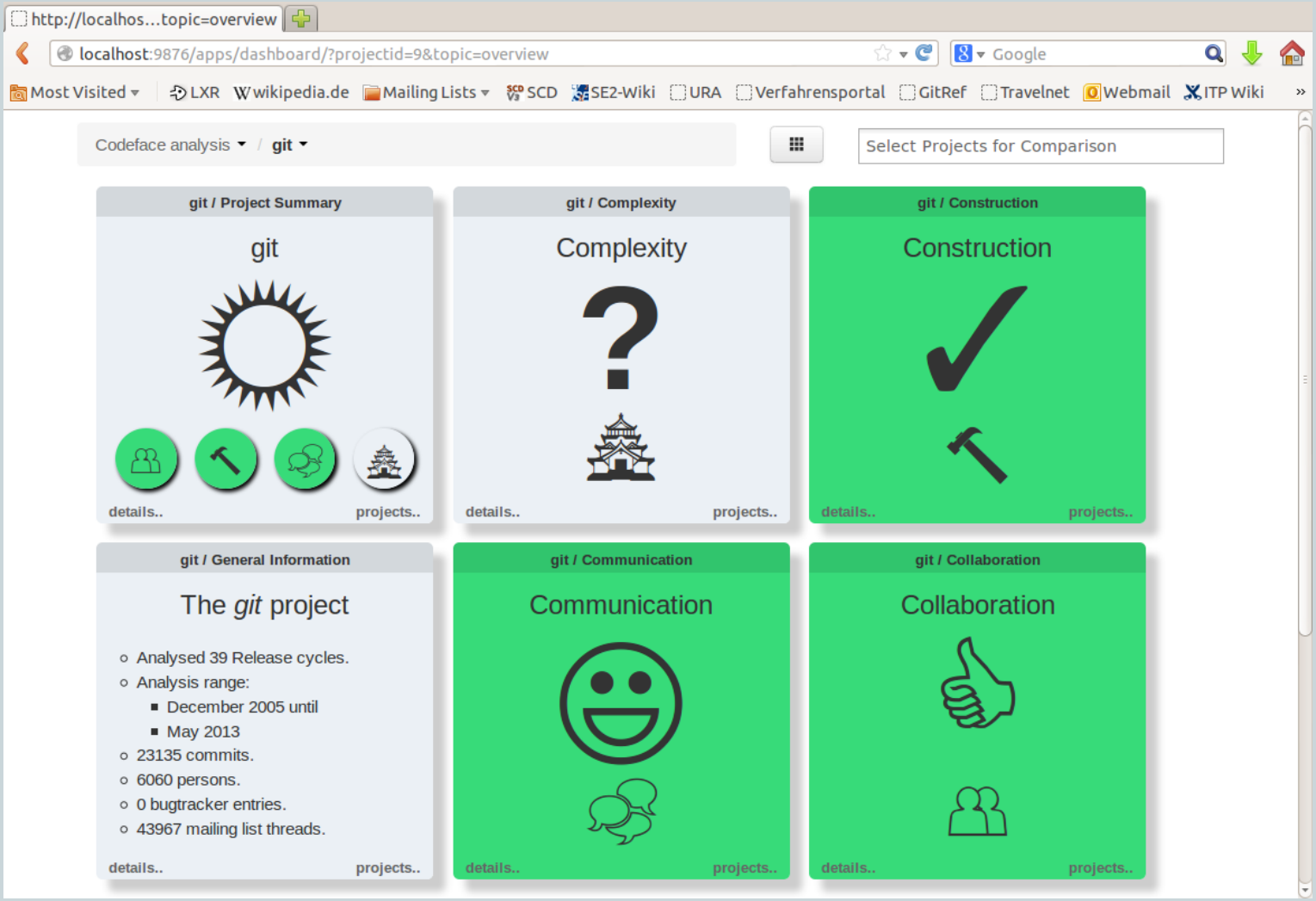
- Qemu, Linux kernel, Apache httpd, Joomla, jQuery, PHP, Busybox, gcc, firefox, Perl, openssl, bootstrap, OpenStack, QT
- See you at the Codeface booth!

Image source: `greensandgills.files.wordpress.com`

Web Frontend

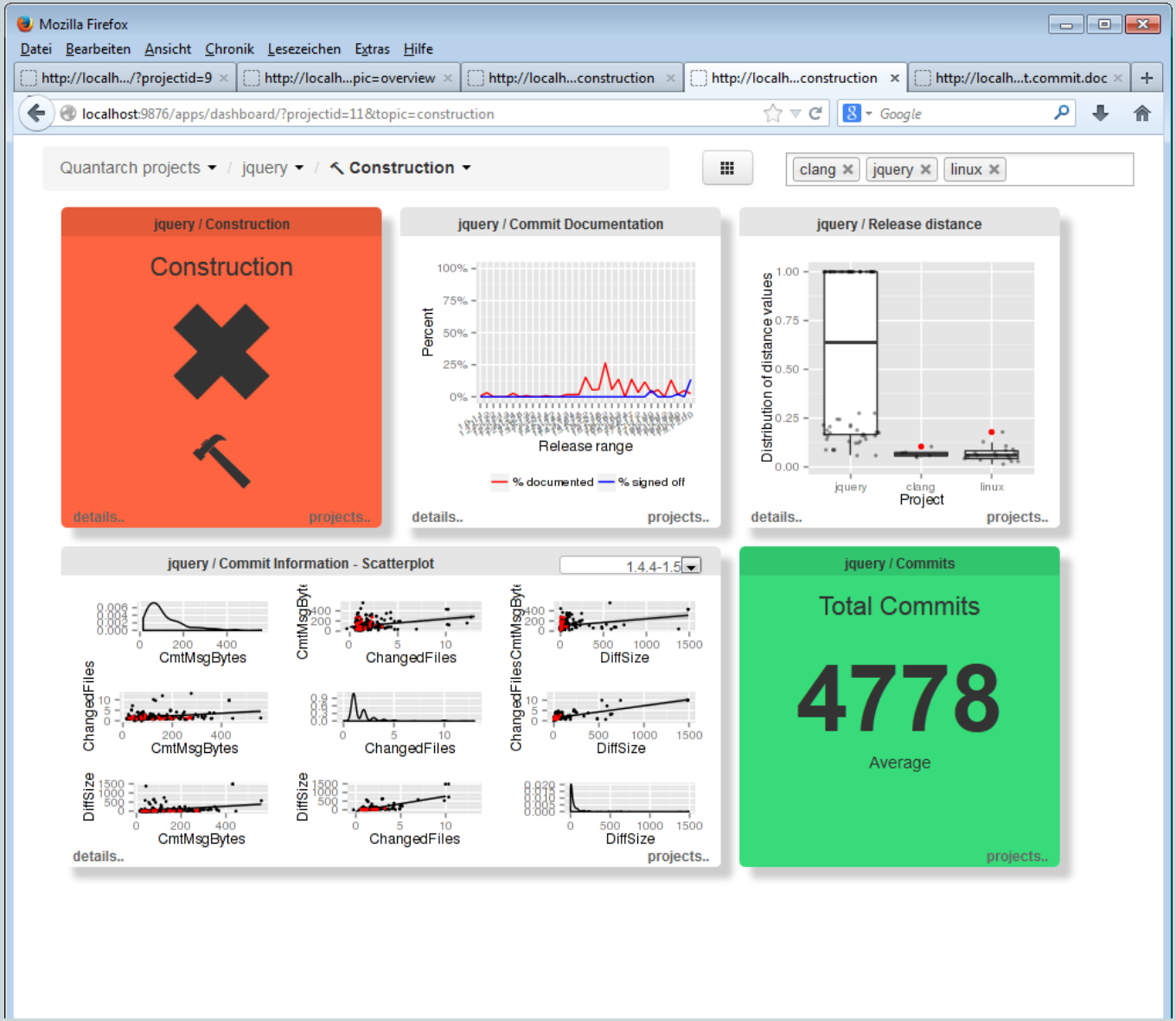
SIEMENS

Web Frontend



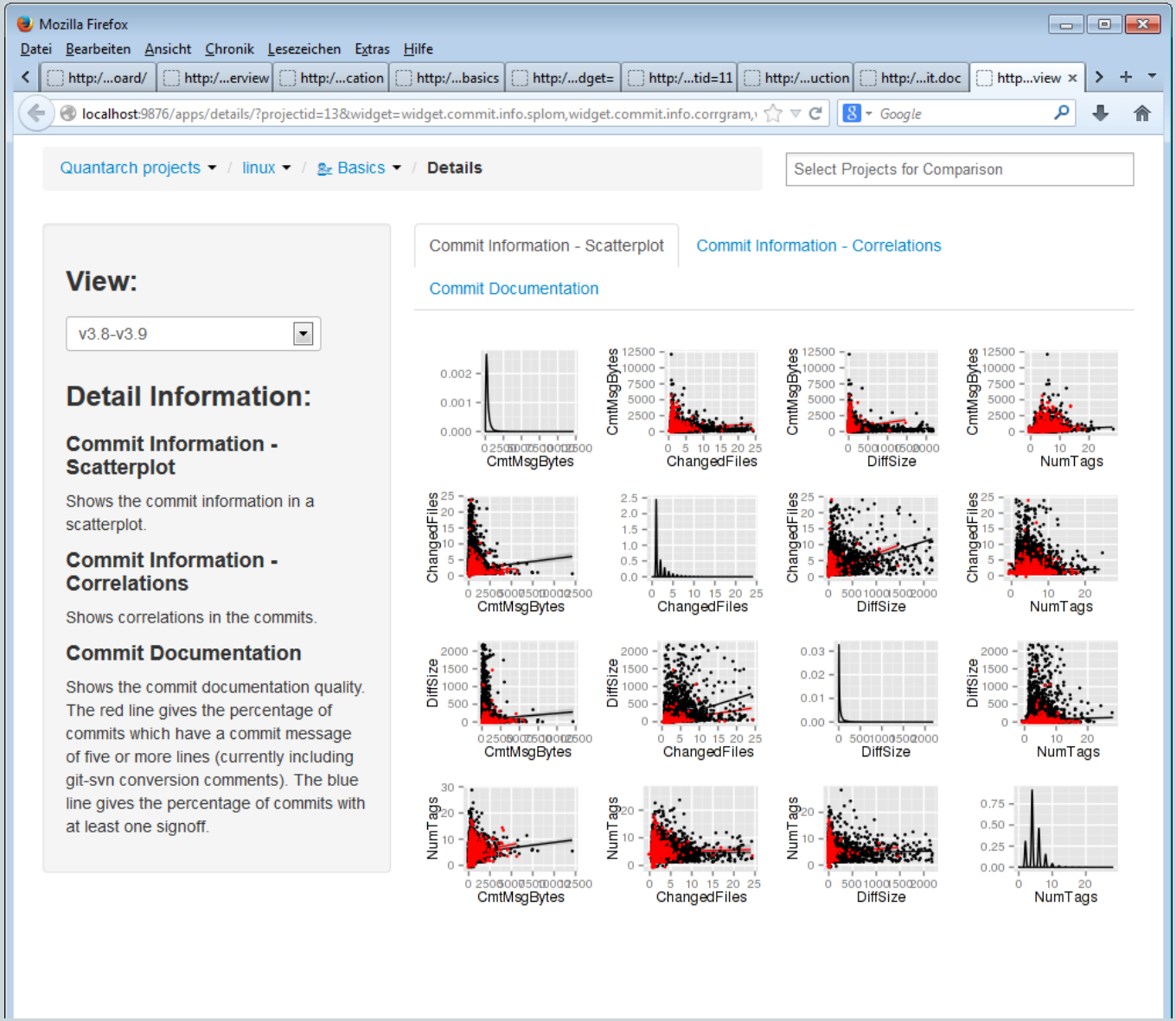
SIEMENS

Web Frontend

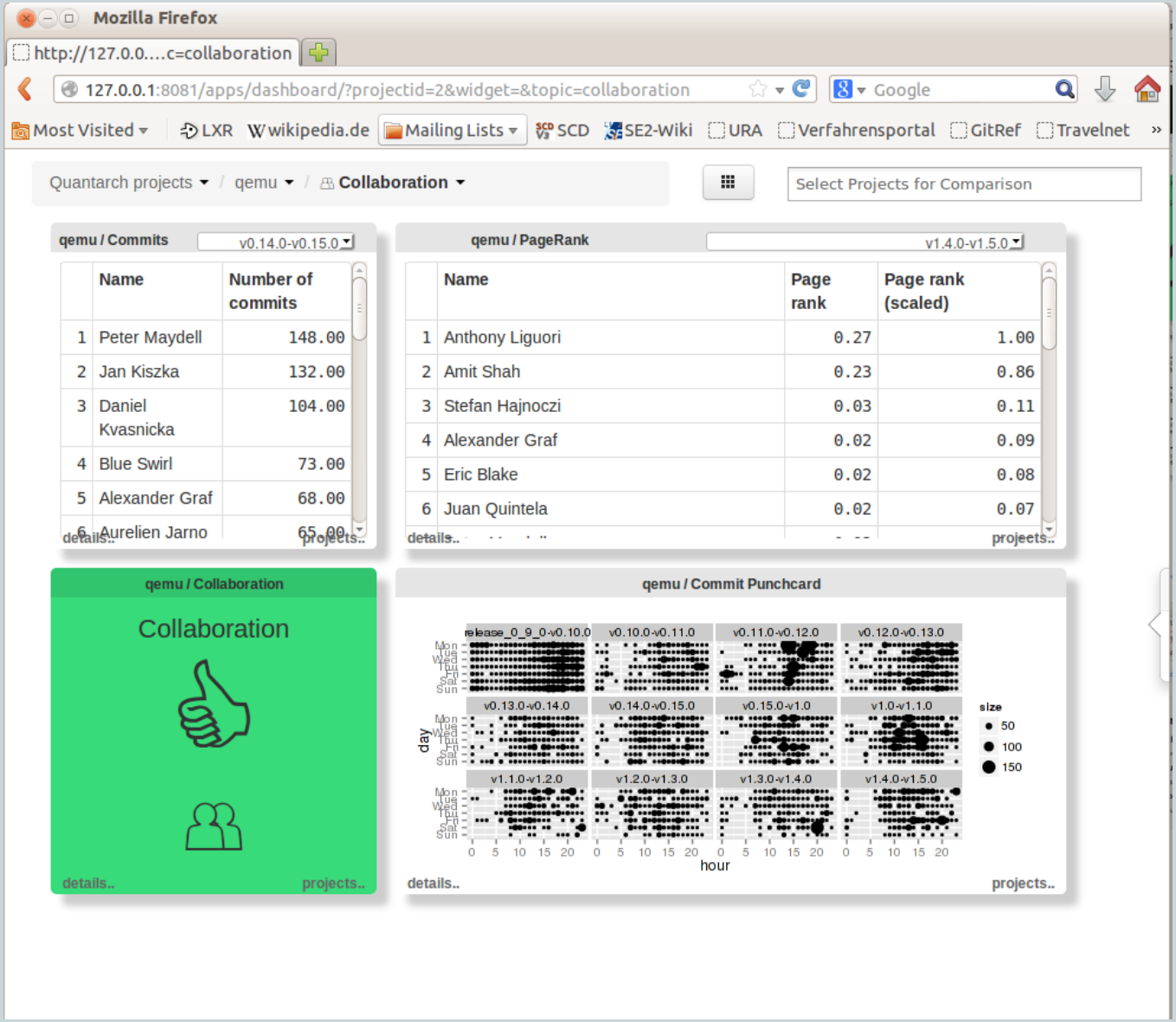


SIEMENS

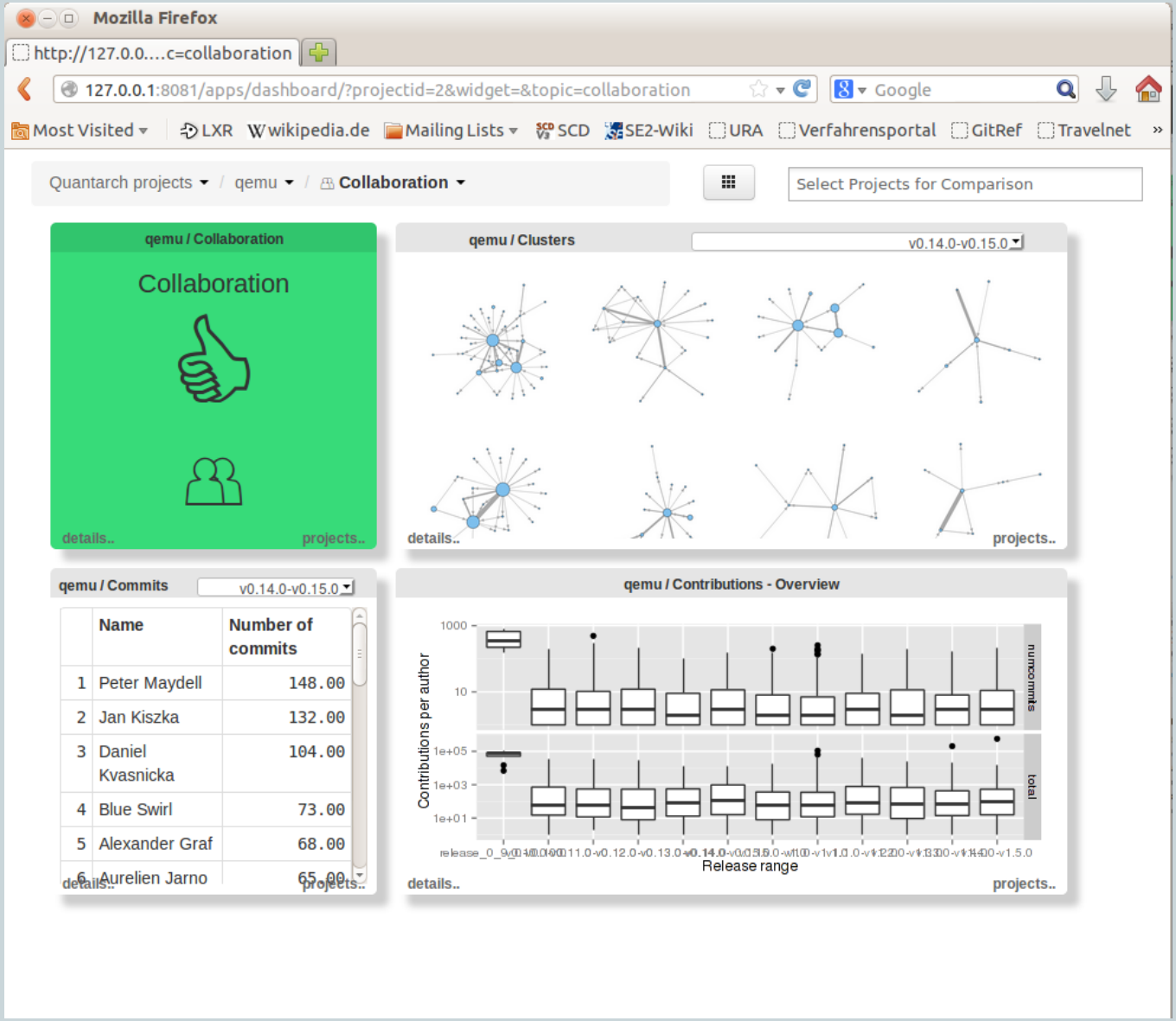
Web Frontend



Web Frontend



Web Frontend



Technological Challenges

Tasks

- Parse Unstructured, real-world data
- Big data handling
- Multivariate, high-dimensional visualisation
- Statistics, Regression, machine learning
- Graph mining, clustering
- Large-scale automation
- Dynamic, reactive web frontend

Technologies

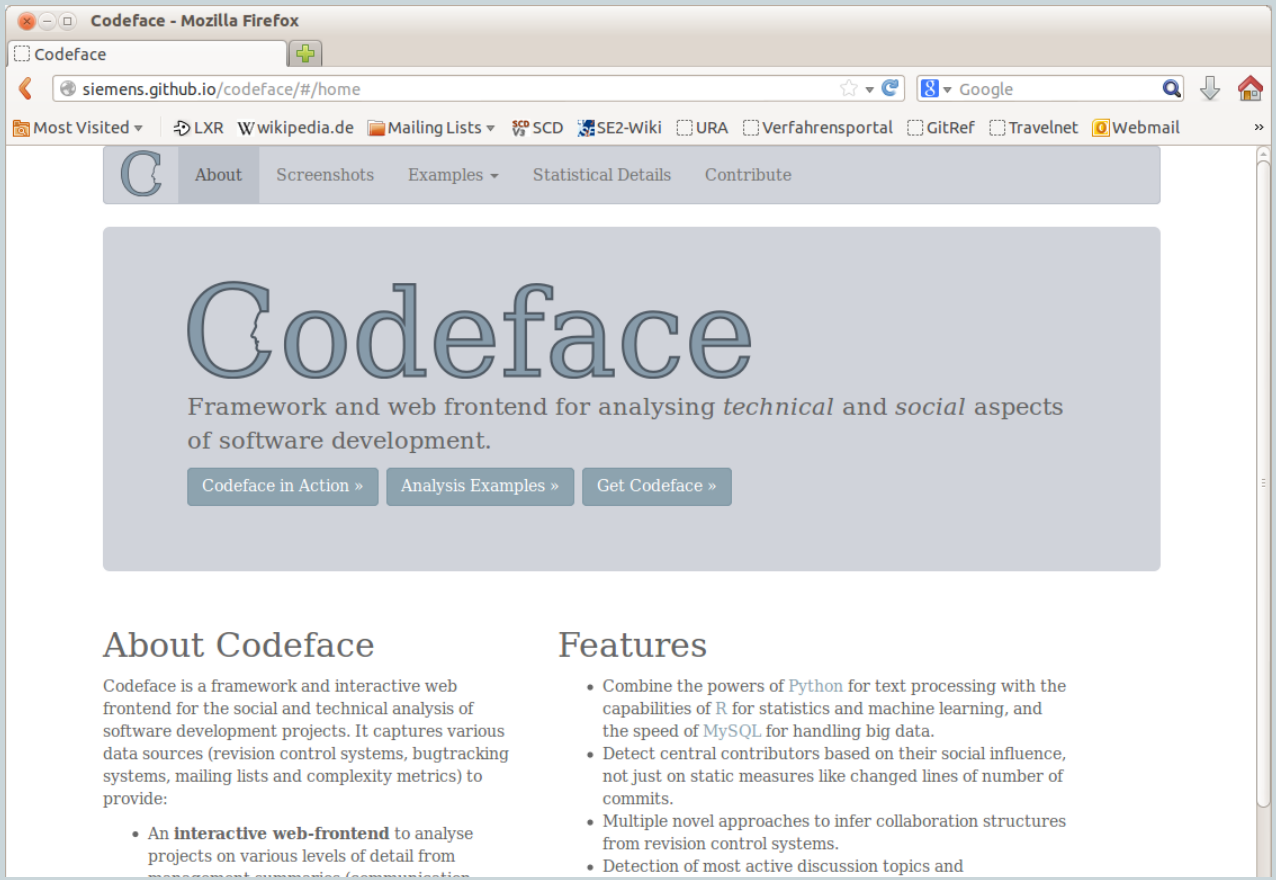
Techniques

Computer Science, Mathematics, Physics, Linguistics, Statistics

Open Source, obviously!

Online Ressources

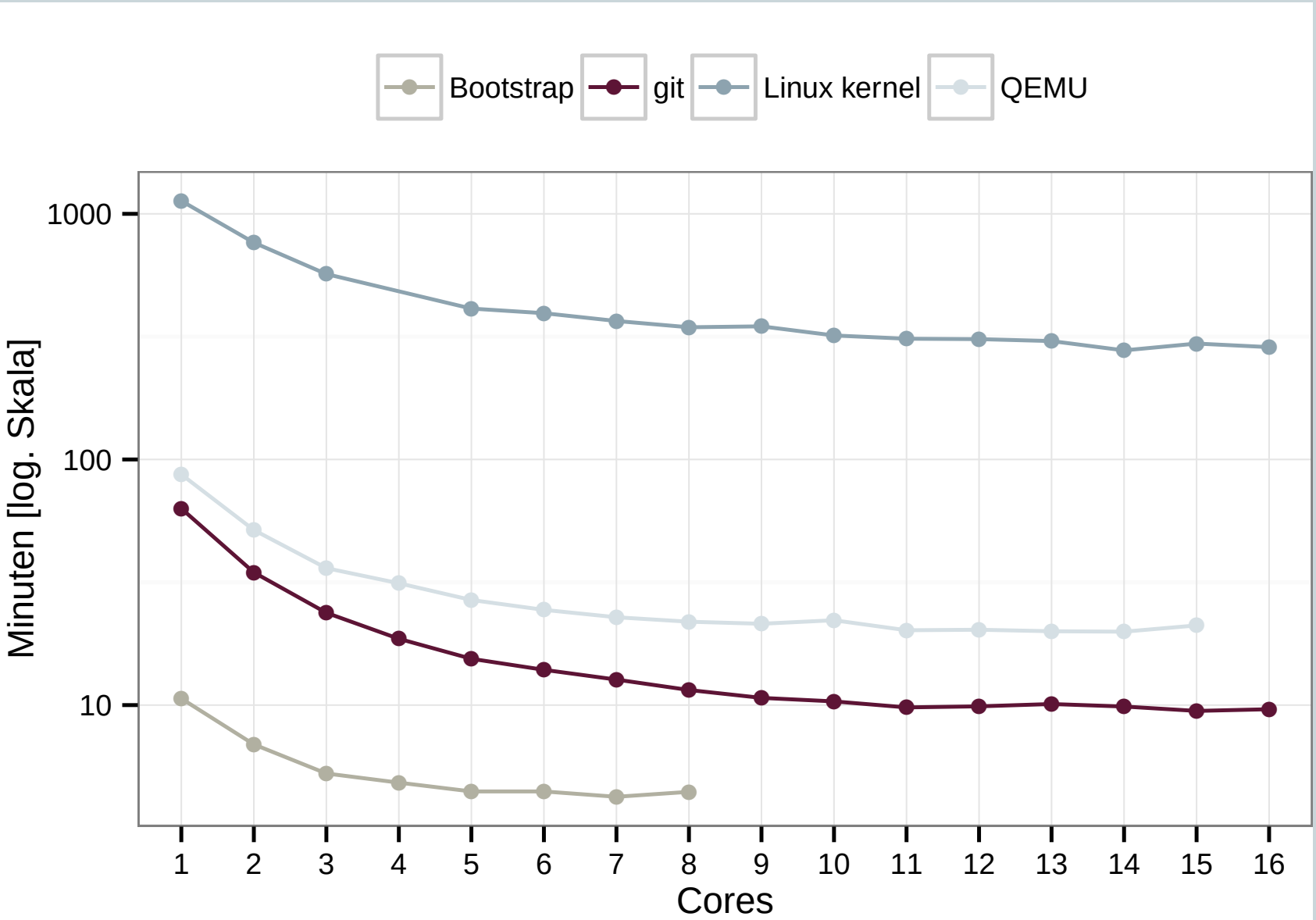
Homepage: `siemens.github.io/codeface`
Github repo: `github.com/siemens/codeface`



Thanks for your interest!

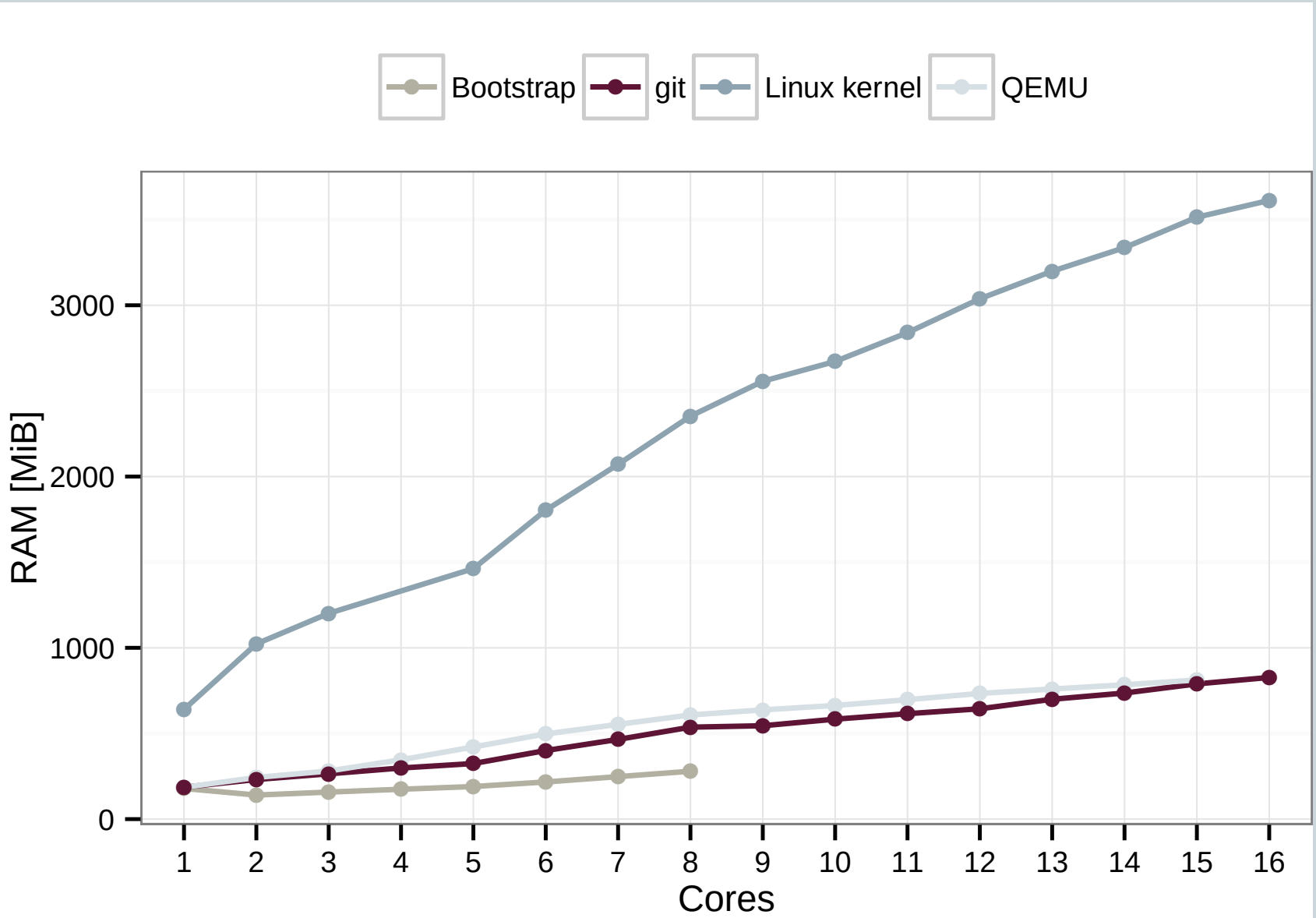
SIEMENS

Backup: Performance Measurements



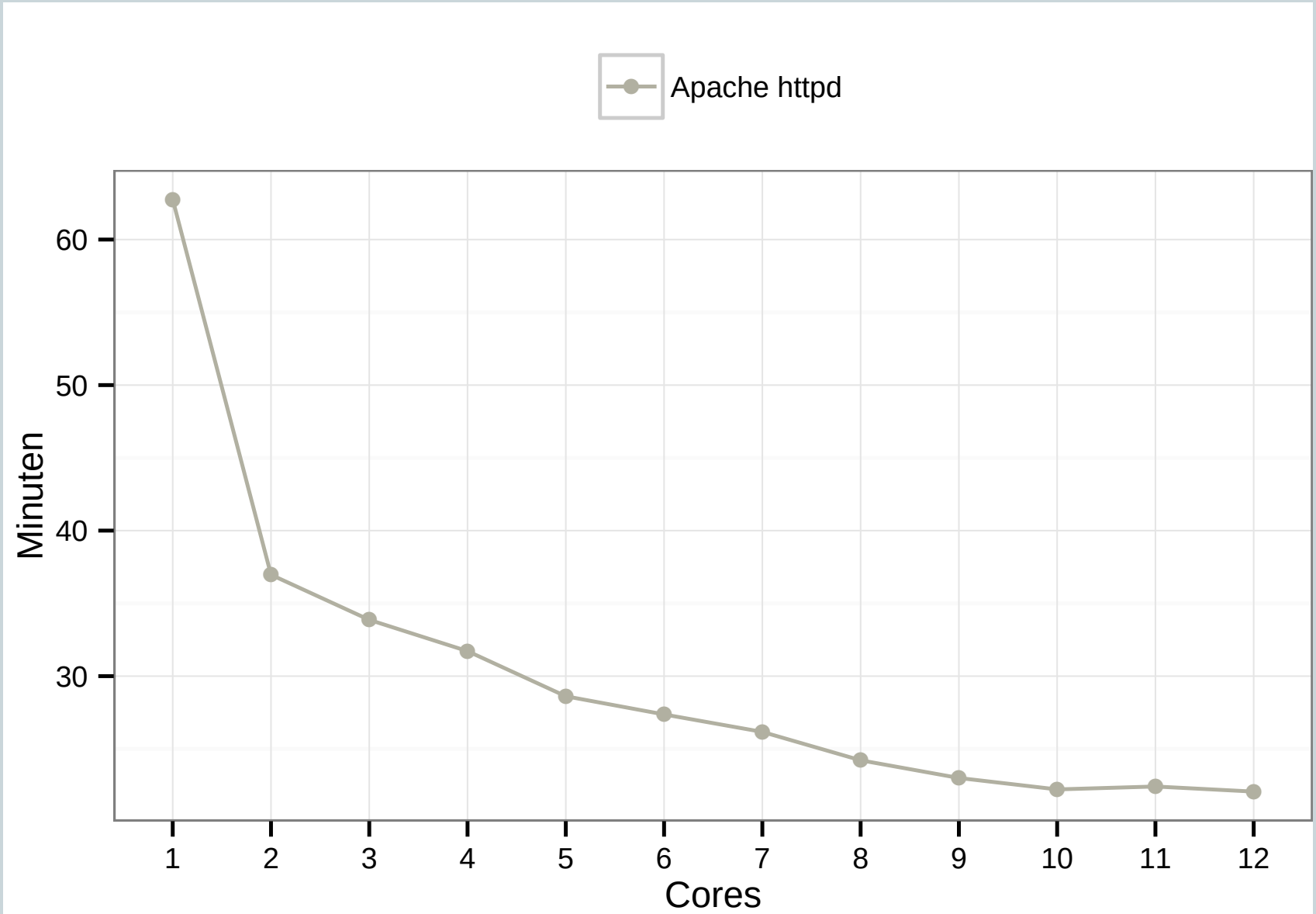
SIEMENS

Backup: Performance Measurements



SIEMENS

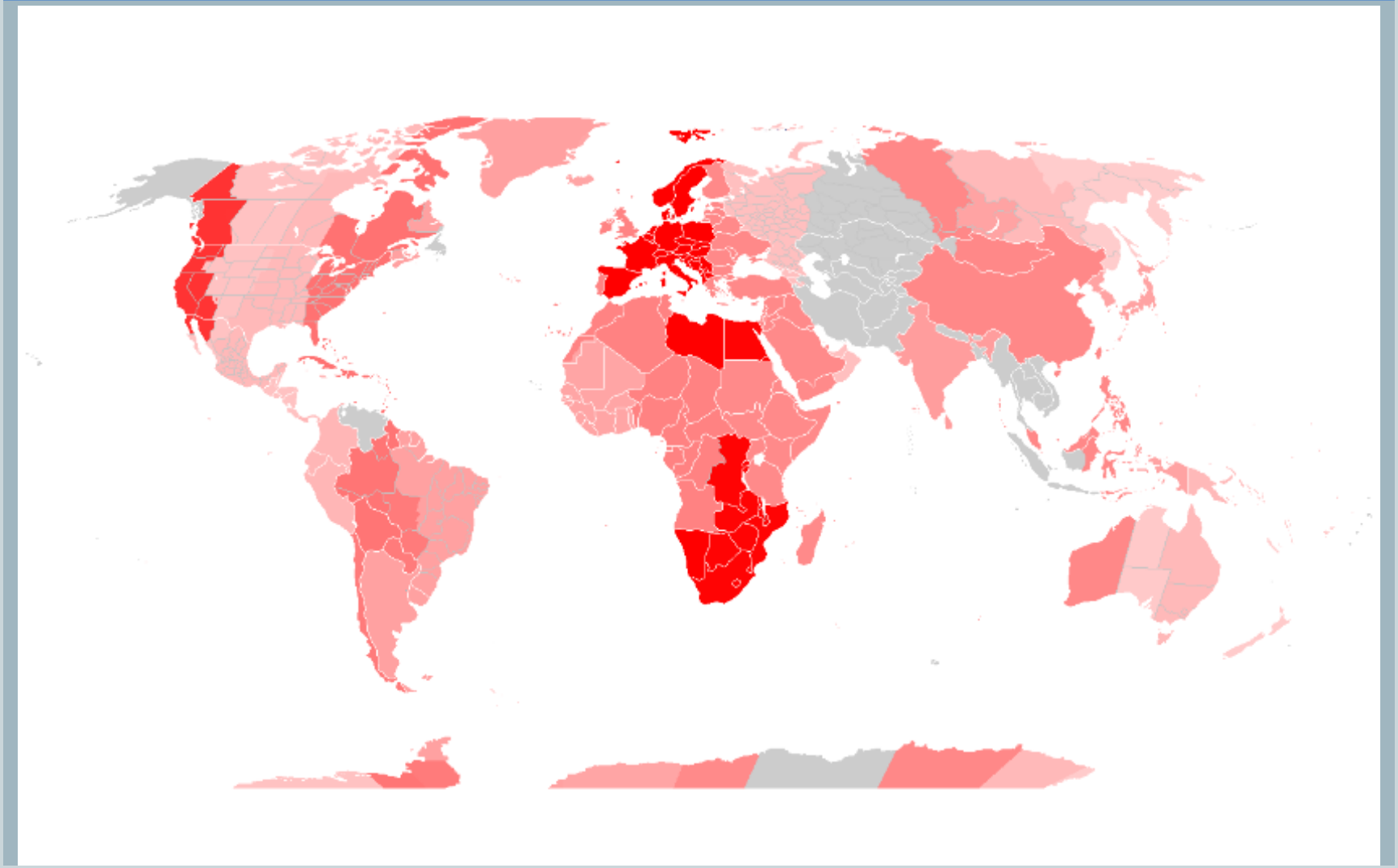
Backup: Performance Measurements



SIEMENS

Backup: Cooperation II

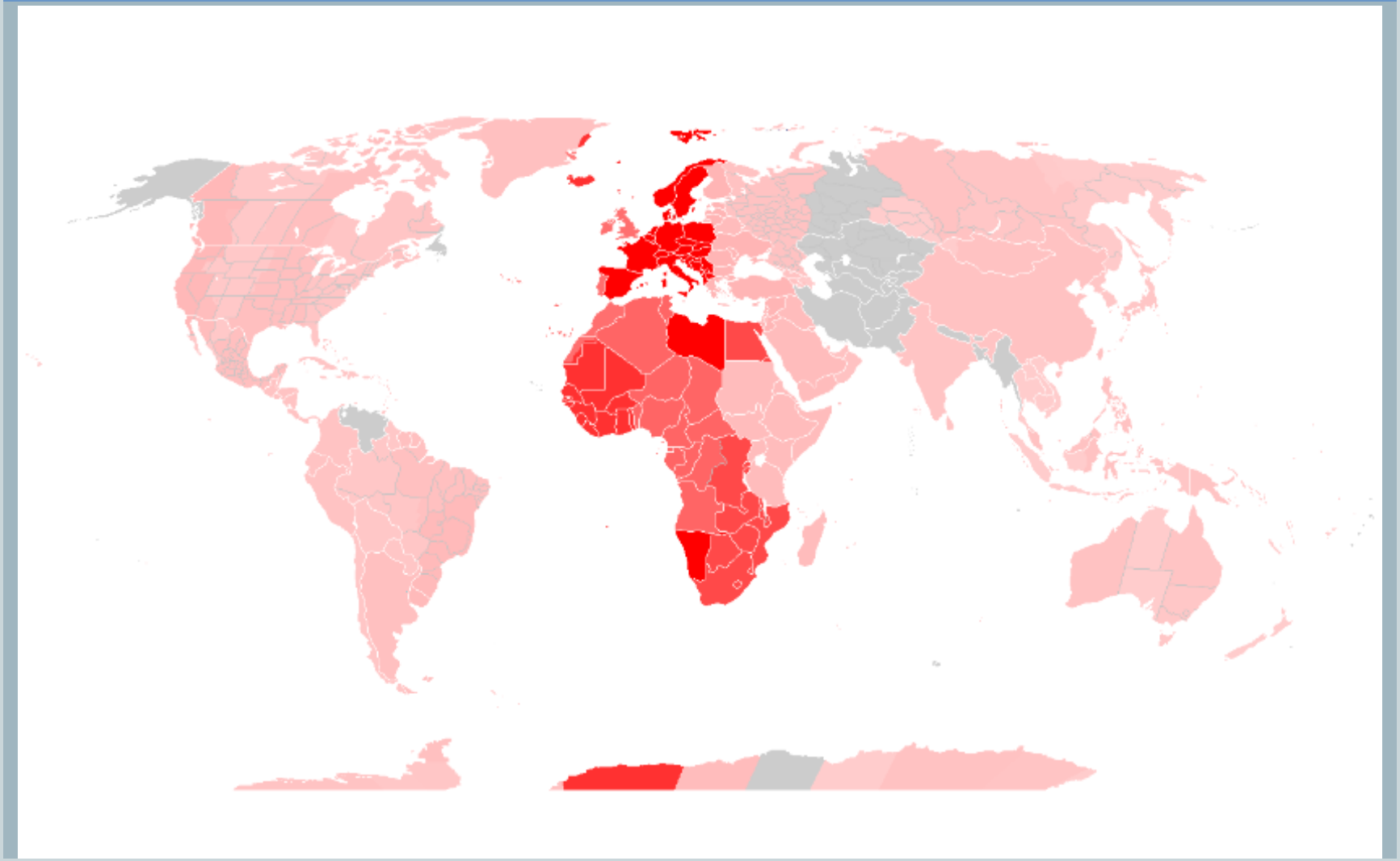
Linux



SIEMENS

Backup: Cooperation II

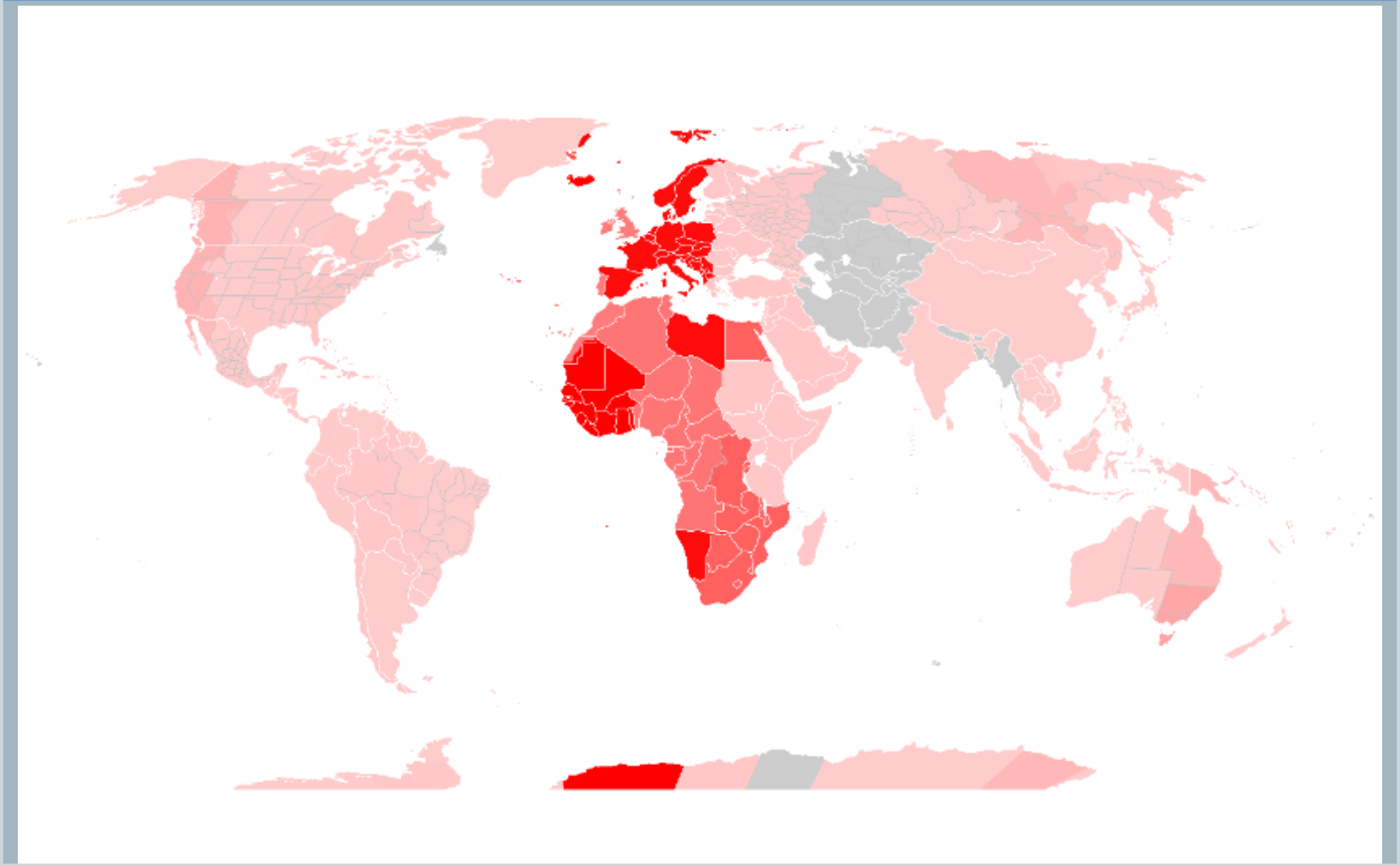
QEMU



SIEMENS

Backup: Cooperation II

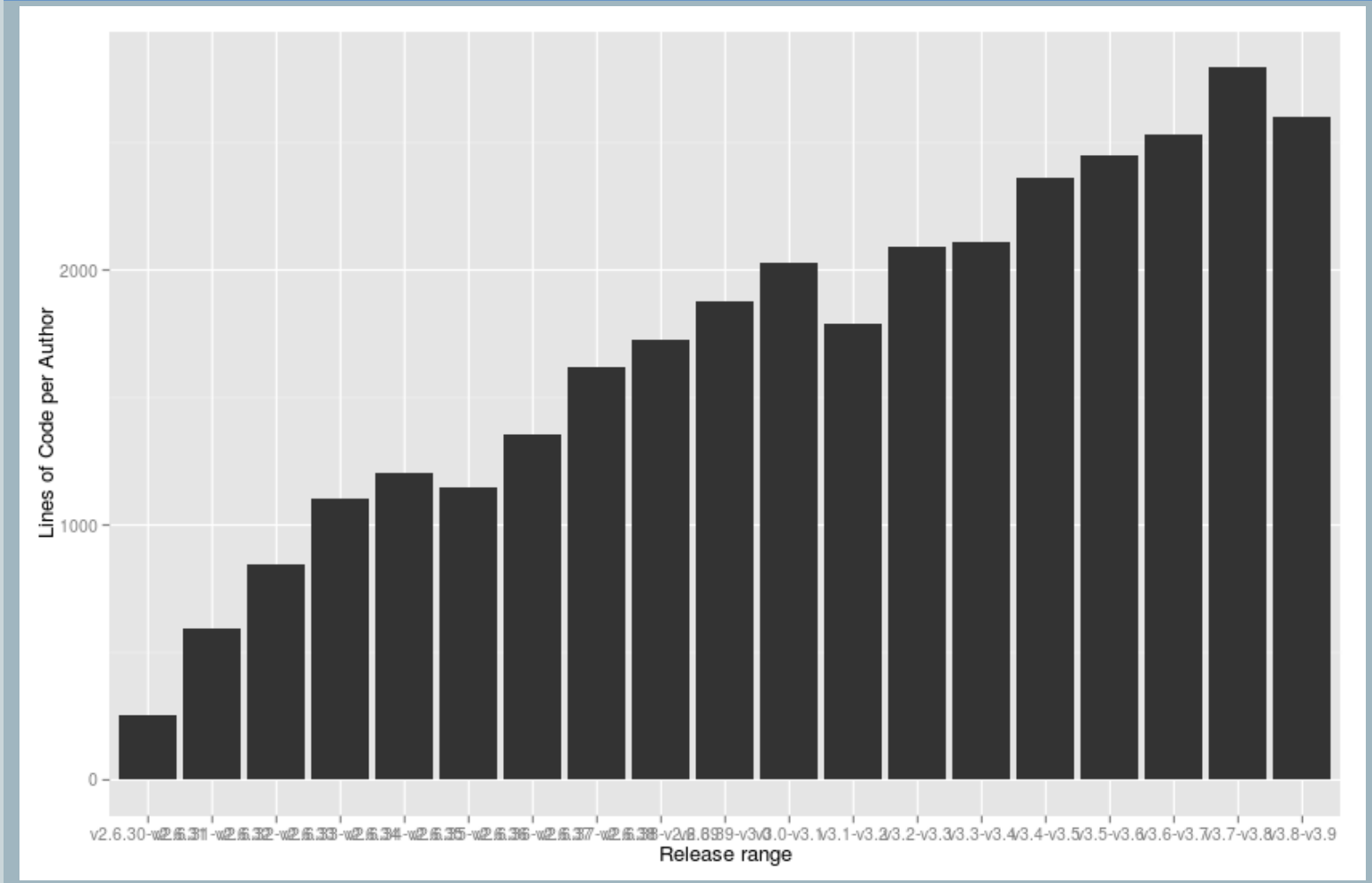
Samba



SIEMENS

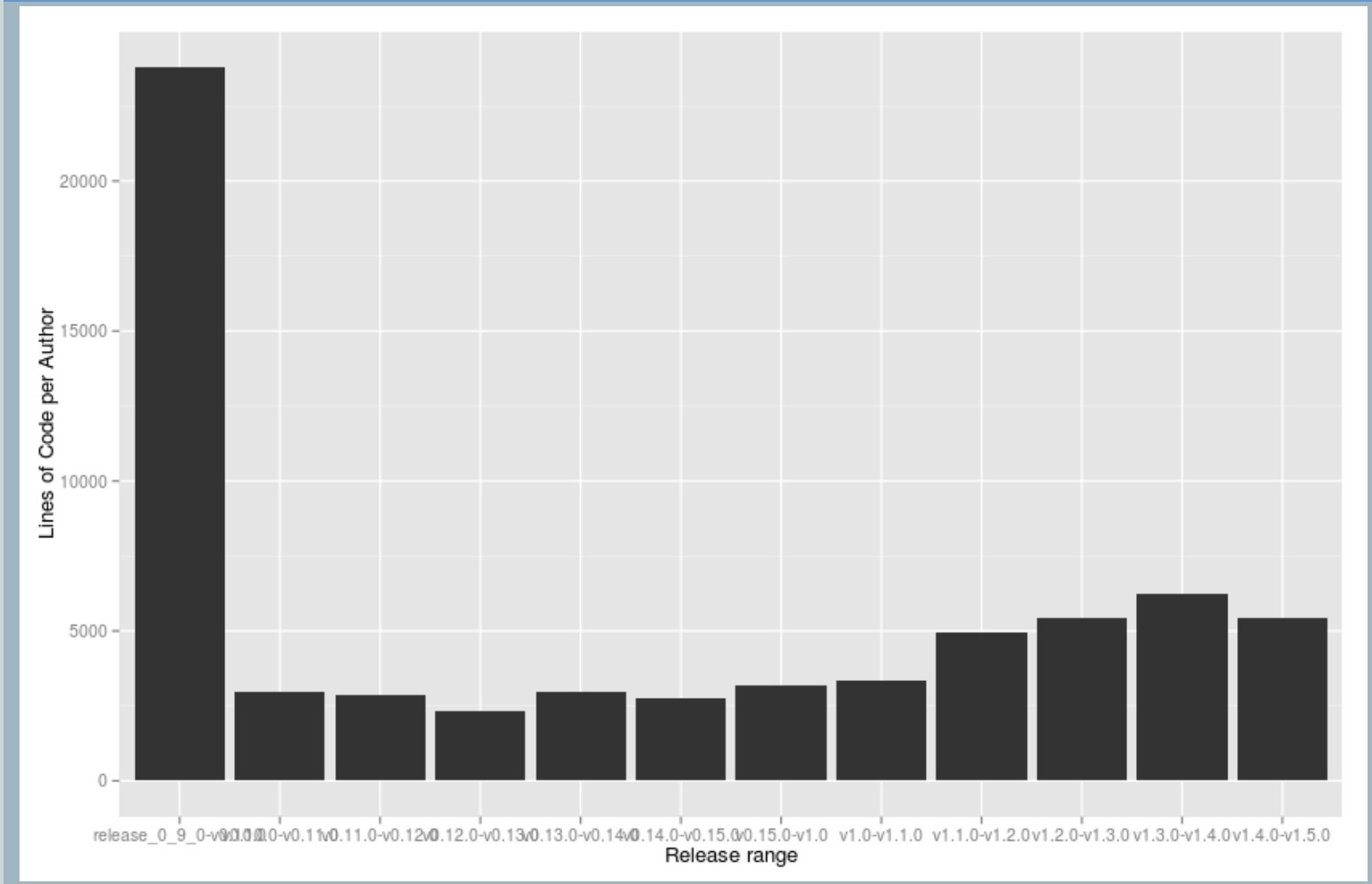
Backup: Maintainance Load

Linux



Backup: Maintainance Load

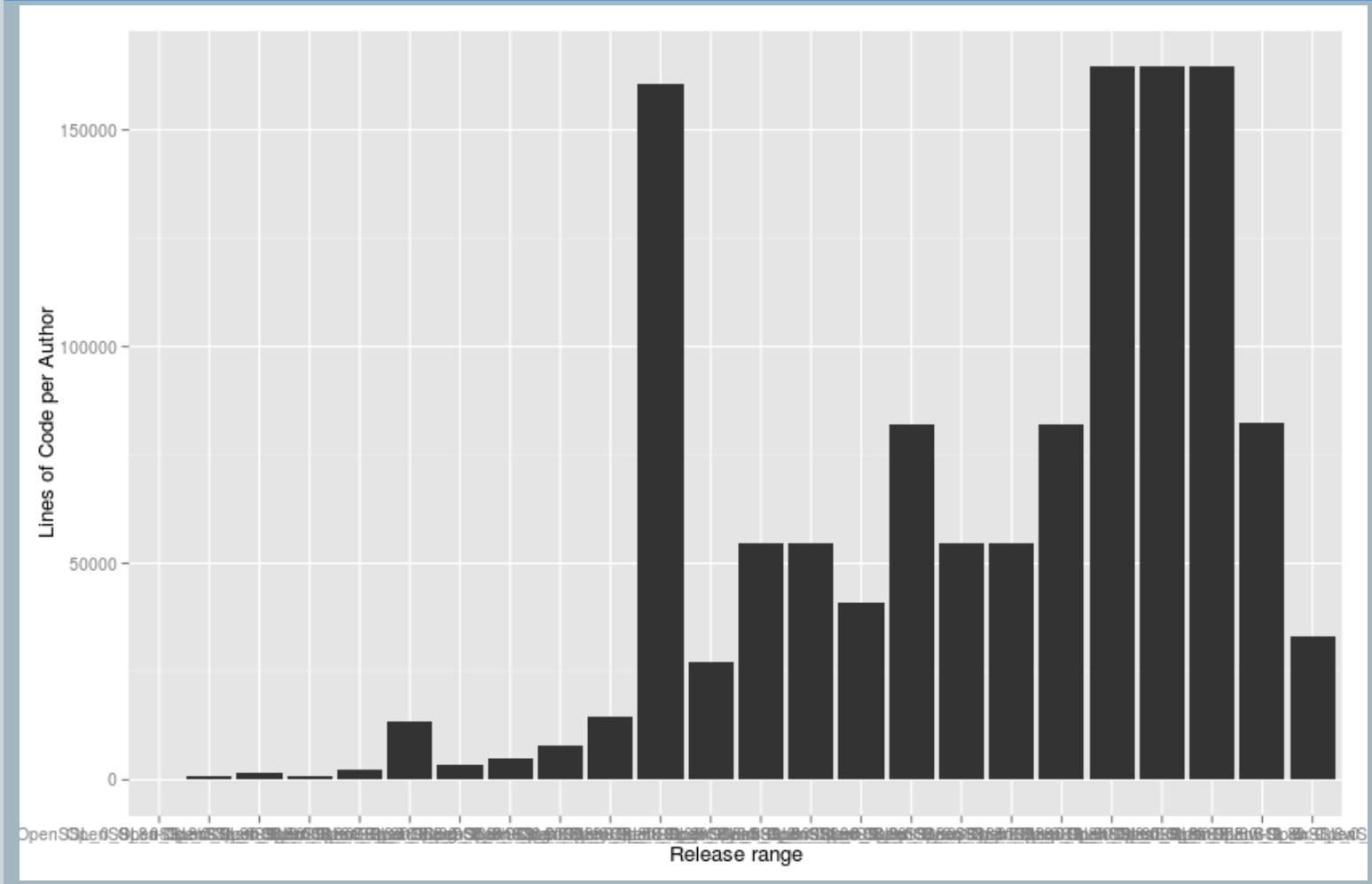
QEMU



SIEMENS

Backup: Maintainance Load

OpenSSL



SIEMENS

Backup: Maintainance Load

Bootstrap